AD-A073 363 DECISIONS AND DESIGNS INC MCLEAN VA F/G 5/10

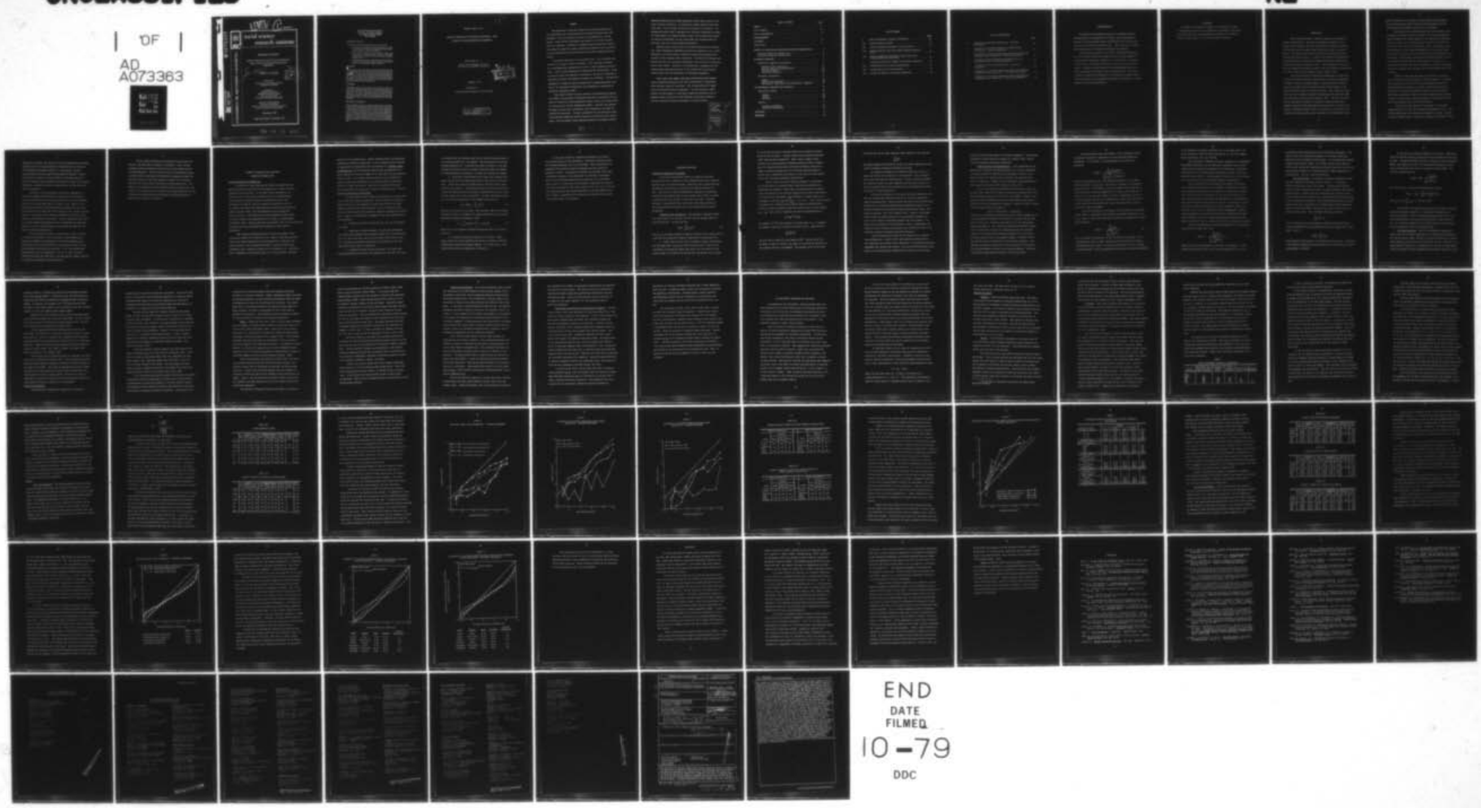
ASSESSING PROBABILITY WITH MULTIPLE INDIVIDUALS: GROUP INTERACT--ETC(U)

DEC 78 D A SEAVER N00014-76-C-0074

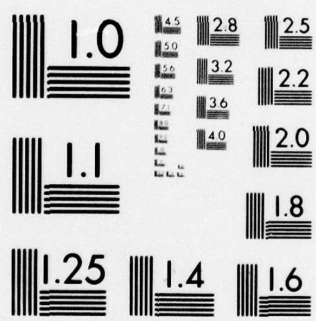
UNCLASSIFIED NL

| OF |

AD
A073363



END
DATE
FILMED
10-79
DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

LEVEL III

AD A073363

DDC FILE COPY

001922-3-T

USC

UNIVERSITY OF SOUTHERN CALIFORNIA

social science research institute

RESEARCH REPORT

Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation

See 1473

DAVID A. SEAVER

Sponsored by:
Advanced Research Projects Agency
Department of Defense

Monitored by:
Decisions & Designs, Inc.
8400 Westpark Drive, Suite 600
McLean, Virginia 22101
Contract No. N00014-76-C-0074, ARPA

December, 1978

SSRI RESEARCH REPORT 78-3

79 08 31 057

Social Science Research Institute
University of Southern California
Los Angeles, California 90007
(213) 741-6955

INSTITUTE GOALS:

The goals of the Social Science Research Institute are threefold:

- To provide an environment in which scientists may pursue their own interests in some blend of basic and methodological research in the investigation of major social problems.
- To provide an environment in which graduate students may receive training in research theory, design and methodology through active participation with senior researchers in ongoing research projects.
- To disseminate information to relevant public and social agencies in order to provide decision makers with the tools and ideas necessary to the formulation of public social policy.

HISTORY:

The Social Science Research Institute, University of Southern California, was established in 1972, with a staff of six. In fiscal year 1978-79, it had a staff of over 90 full- and part-time researchers and support personnel. SSRI draws upon most University academic Departments and Schools to make up its research staff, e.g. Industrial and Systems Engineering, the Law School, Psychology, Public Administration, Safety and Systems Management, and others. Senior researchers have joint appointments and most actively combine research with teaching.

FUNDING:

SSRI Reports directly to the Executive Vice President of USC. It is provided with modest annual basic support for administration, operations, and program development. The major sources of funding support are federal, state, and local funding agencies and private foundations and organizations. The list of sponsors has recently expanded to include governments outside the United States. Total funding has increased from approximately $150,000 in 1972 to almost $3,000,000 in the fiscal year 1978-1979.

RESEARCH INTERESTS:

Each senior SSRI scientist is encouraged to pursue his or her own research interests, subject to availability of funding. These interests are diverse; a recent count identified 27. Four major interests persist among groups of SSRI researchers: crime control and criminal justice, methods of dispute resolution and alternatives to the courts, use of administration records for demographic and other research purposes, and exploitation of applications of decision analysis to public decision making and program evaluation. But many SSRI projects do not fall into these categories. Most projects combine the skills of several scientists, often from different disciplines. As SSRI research personnel change, its interests will change also.

1

Research Report 78-3

ASSESSING PROBABILITY WITH MULTIPLE INDIVIDUALS: GROUP INTERACTION VERSUS MATHEMATICAL AGGREGATION

David Arden Seaver

Social Science Research Institute
University of Southern California

December, 1978

Sponsored by

Defense Advanced Research Projects Agency

SUMMARY

The application of decision theory often involves assessing subjective probabilities, and procedures for assessing them are quite well developed. But such procedures are based on assessments by a single person. Often multiple individuals are called on to provide the probabilistic judgments. Unanimity in judgments among the multiple individuals cannot be expected, thereby creating the problem of how to arrive at a single probability distribution that can be used in applying decision theory.

Two general approaches to this problem exist. The individuals can interact as a group to reach a consensus, or the individual judgments can be mathematically aggregated to produce a single probability distribution. Each of these approaches has advantages and disadvantages. Group interaction allows the exchange of information, but may be susceptible to dominance by certain individuals or pressure for conformity. Mathematical aggregation is simple to use and ensures that a single distribution will result, but theoretical difficulties are encountered in specifying an appropriate aggregation model.

Using several forms of group interaction and mathematical aggregation models, this research investigated the quality of probabilities produced by interaction versus mathematical models, and by the various forms of interaction and various mathematical models. "Quality" was measured by proper scoring rules, calibration, and extremeness on two types of probability assessments: discrete assessments for two-alternative questions and beta probability density functions for questions about percentages. Ten four-person groups comprised primarily of graduate students

assessed probabilities for twenty questions of each type in each of five types of group interaction: no interaction, Delphi, Nominal Group Technique (NGT), a mix of Delphi and NGT, and discussion to consensus. The mathematical models used to aggregate the individual assessments included the linear model, the weighted geometric mean, and the pari-mutuel model for discrete assessments; and the linear model and conjugate model for densities; each with various weighting procedures.

Applying proper scoring rules to the group probabilities indicated that simple mathematical aggregation without any interaction, e.g. linear aggregation with equal weights, generally produced group probabilities as good as those assessed after interaction. Interaction did produce more extreme but less well calibrated assessments, with the type of interaction having little effect. Generally, the calibration of mathematically aggregated group probabilities prior to any interaction was quite good, clearly better than the calibration of individual assessments.

These results may appear relatively uninteresting from a psychological perspective because of the lack of differences in assessments after different types of interaction. But the implications for applications of decision theory are important. In many instances, simple, mathematical aggregation of individual probability assessments may be adequate without resorting to more elaborate, practically difficult, and time consuming interactive processes or modeling efforts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ACKNOWLEDGEMENTS

This research was supported in part by the Advanced Research Projects Agency of the Department of Defense, and was monitored by the Office of Naval Research under contract N00014-76-C-0074 and subcontract with Decisions and Designs, Inc.; and in part by the Sonderforschungsbereich 24, University of Mannheim, West Germany.

I am especially grateful to my advisor, Professor Ward Edwards, for his continued guidance and support, and to the other members of my dissertation committee, Professors J. E. Keith Smith, co-chairman, J. Frank Yates, and William A. Ericson, for their assistance and patience. I would also like to thank Katrin Borcherding and the research assistants at the SFB for their invaluable help. And finally, I thank my friends and colleagues at the University of Southern California for providing the stimulating and supportive atmosphere in which research can best be accomplished.

DISCLAIMER

The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or of the United States Government.

INTRODUCTION

One of the cornerstones of decision theory is the concept of subjective probability. The theory of subjective probability (e.g., Savage, 1954) provides a basis for quantifying the subjective opinions of a decision maker or experts whose opinions are used by a decision maker in the probabilistic terms which can then be used explicitly in the decision making process. In order to use subjective probabilities, techniques have been developed for assessing subjective probabilities (Spetzler and Stael von Holstein, 1975). The development of the theory and the assessment techniques has led to the use of subjective probability in a wide variety of real decision contexts (Beach, 1976).

But the applications of subjective probabilities in real-world contexts have also illuminated a gap between the theory and assessment techniques, and the technology needed in certain decision situations. Often groups rather than individuals are the decision makers or the experts providing input to the decision makers. And research has shown that the type of judgments required are generally more valid when made by groups rather than individuals (Seaver, 1976). Yet both the theory and the assessment techniques of subjective probability have been primarily oriented toward quantifying the uncertainty of a single individual. Although as Savage (1954, p. 8) points out, the theory is not limited to the single person case, extensions to the multi-person case

depend on some sort of unanimity of action among the group members. Such unanimity rarely exists in decision making groups until some process specifically aimed at achieving it is undertaken.

One possible way in which to create a form of unanimity is for the group to interact to reach a consensus. But social pychological research suggests that several aspects of the interaction process may reduce the quality of the resulting consensus (Collins and Guetzkow, 1964; Davis, 1969; Van de Ven and Delbecq, 1971). For example, interacting groups will often expend considerable time and effort simply structuring the group and the interacting process, both explicitly and unknowingly. Additionally, dominance by individuals because of status or personality may decrease the effectiveness of the group. Or, pressure for conformity may cause the group to, in effect, make simply reaching an agreement more important than the substantive value of the consensus.

Elaborate interactive processes that attempt to circumvent these factors have been the subject of extensive research. Typically, such processes rely on strictly controlled interaction and do not actually produce a consensus, but rather necessitate some type of aggregation of individual judgments to produce the group judgment. Since these processes are often quite time-consuming and their effectiveness is questionable, simpler approaches to the problem of determining group probabilities should be considered.

One obvious simple approach is to average the individual probabilities; or use some other mathematical aggregation rule. Theoretical difficulties with mathematical aggregation do exist, however, as shown by Dalkey (1972). He proved, in the spirit of Arrow's (1951)

Impossibility Theorem, that there is no rule for aggregating individual probabilities into a group probability distribution that satisfies a set of seemingly reasonable conditions. Additionally, the more rigorous and theoretically appealing mathematical aggregation models are difficult to apply in practice because an inordinate amount of data or extremely complex judgments are required as inputs to the models. Simplifying, although unrealistic, assumptions can be made that allow use of these models.

Although it has some problems, mathematical aggregation of individual probabilities does have two advantages over interaction: the group probability will always be produced, and it will be obtained using less of the decision makers' or experts' time. Whether or not mathematical aggregation should generally be advocated for obtaining group probabilities should and/or would depend on two, probably related, factors: the quality of the resulting probabilities, and the acceptability of the procedure to the group. In fact, should the group agree to use some mathematical aggregation rule to determine the group probabilities, it is in effect producing the unanimity necessary for the theory of subjective probability.

Thus, the question of what is the best way to reach unanimity is an empirical question. Will the quality of group probabilities produced by mathematical aggregation of individual probabilities be good enough so that such a procedure can be advocated rather than the much more cumbersome interaction processes? If so, what mathematical model should be used for aggregation? If not, is there a specific interactive process that works best? The experimental research reported here explores the answers to these questions.

However, before describing the experiment and presenting the results, some additional information is presented. First, several concepts concerning probability and its use in this research are defined and explained. Then the specific nature of the different types of both interaction processes and mathematical aggregation models are described, along with the scant empirical research on the relative merits of the various means of determining group probabilities. Subsequently, the experiment and the obtained results are presented. And, finally, the implications of the research for groups faced with the task of determining probabilities are discussed with special emphasis on applications in realistic situations.

CONCEPTS IN ASSESSING AND EVALUATING SUBJECTIVE PROBABILITIES

Assessing Subjective Probabilities

Procedures for both assessing and evaluating subjective probabilities depend on the nature of the propositions or events for which probabilities are assessed. If the events under consideration are discrete--that is, the space of possible events is represented by a finite number of mutually exclusive and exhaustive events--then assessments can take the form of a probability between 0.0 and 1.0. If, however, the events are represented by a continuum with an infinite number of possibilities, then the assessments must be probability density functions. Procedures for eliciting probability density functions often produce only approximations (cf. Seaver, von Winterfeldt, and Edwards, 1978). Spetzler and Stael von Holstein (1975) discuss particular procedures for eliciting appropriate judgments for both types of assessments.

When complete probability density functions are needed, often a particular family of distributions (e.g., normal or beta distributions) can provide enough flexibility by varying parameters to represent subjective opinion. This is especially useful in certain instances when information from a variety of sources is to be combined; e.g., subjective prior probability with objective data, or, in some instances, multiple

subjective prior probabilities. Bayes' Theorem provides the appropriate mechanism for combining information. If the information being combined is represented by distributions that are members of a _conjugate family of distributions_, the distribution resulting from the application of Bayes' Theorem will also be a member of the same family of distributions (DeGroot, 1970). For example, beta distributions can be combined to produce another beta distribution, or combining normal distributions produces a normal distribution. And use of conjugate distributions greatly simplifies the computation necessary in applying Bayes' Theorem.

Evaluating Subjective Probabilities

In a philosophical sense, subjective probabilities by their very nature cannot be externally evaluated. They are judgments or opinions, and as such can only be evaluated in terms of how well the elicited judgment represents the internal opinion. But in a practical sense, certain criteria characterize properties subjective probabilities should have. Seaver, von Winterfeldt, and Edwards (1978) have identified five such desiderata:

1. Assessments should be consistent with the laws of probability theory.

2. Assessments should be extreme. For discrete assessments, this implies that probabilities assigned to events that occur should be near 1.0, while those assigned to non-occurring events should be near 0.0. Continuous assessments should have a high density at the true value and a density near 0.0 elsewhere.

3. Assessments should be well-calibrated. This means that multiple assessments should have the property that the events for which

the probabilities are assessed occur with a relative frequency equal to the assessed probability. For example, discrete events for which the assessed probability is .75 should occur about 75 percent of the time. And about 50 percent of the true values should fall below the medians of assessed probability densities, or within the interquartile ranges.

4. Assessments should produce high scores when evaluated with proper scoring rules (see Murphy and Winkler, 1970; Stael von Holstein, 1971). These scores measure a combination of criteria 2 and 3, which typically will conflict. The defining property of proper scoring rules is that the expected value of the score is maximized if and only if the assessor reports his or her true opinion. An often used proper scoring rule for discrete assessments is the quadratic scoring rule:

$$S_k = 2p(\Theta_k) - \sum_{j=1}^{n} p(\Theta_j)^2 \quad (1)$$

where S_k is the score if Θ_k occurs. The continuous form of the ranked probability score is an example of a proper scoring rule for continuous assessments (Matheson and Winkler, 1976):

$$S = -\int_{-\infty}^{t} P(\Theta)d\Theta - \int_{t}^{\infty} (1 - P(\Theta))^2 d\Theta \quad (2)$$

where $P(\Theta)$ is the cumulative assessed distribution and t is the true value of Θ.

5. Assessments should be responsive to evidence. Seaver et al. suggest this means probabilities should be revised as evidence accumulates as specified by Bayes' Theorem. In a formal sense, this follows from the laws of probability (criterion 1.).

In any given situation, probability assessments are usually not evaluated using all five desiderata. Elicitation procedures often do not allow properties 1 or 5 to be violated. Most investigations of procedures for assessing subjective probabilities have focused on properties 3 and 4. Lichtenstein, Fischhoff, and Phillips (1977) have reviewed the research on the calibration of (individual) assessments, most of which indicated assessments are usually not well-calibrated. Scores tend to vary depending on the assessor's expertise and training (cf. Stael von Holstein, 1971, 1972; Winkler, 1971), but often scores are only slightly better than would be achieved with uniform probabilities. Thus, clearly, assessments can be improved, and using multiple persons is a possible means of improvement.

ASSESSMENT APPROACHES

Mathematical Aggregation Procedures

A variety of mathematical models for combining individual probabilities into a composite or group probability have been suggested. Depending upon the particular model, these models may be applicable for aggregating either discrete probabilities or density functions, or both. Some are quite simple mathematically, although the underlying theoretical justification may be quite complex; while others are quite complicated and often unusable in realistic situations. Although unusable in their general form, still these more complex models are practically beneficial because they can be simplified with certain assumptions.

Weighted linear combination. This procedure, sometimes called the "opinion pool," can be used with both discrete probabilities and density functions. It takes the form

$$p_G(\Theta) = \sum_{i=1}^{m} w_i p_i(\Theta) \tag{3}$$

where p_G is the group probability (density function) and w_i and p_i are the weight and the probability (density) respectively of individuals $i=1, \ldots, m$. Stone (1961) was the first to present a formal justification for this model when, assuming a convex utility function common to all individuals, he proved the rather weak result that the utility of the decision made on the basis of an opinion pool was greater than or equal

to the minimum utility of a decision based on the probability distribution of any individual. Stronger results were obtained by Bacharach (1975) using stronger assumptions. Again, given a common utility function, and a group preference ordering satisfying forms of independence of irrelevant alternatives and Pareto Optimality, along with a couple of technical assumptions; then the group maximizes expected utility given a probability distribution in the form of linear combination of the individual probability distributions.

DeGroot (1974) has taken a different approach to formalizing the justification for weighted linear combination of probabilities. Individuals are assumed to revise their own probabilities as weighted linear combinations of the revealed probabilities of other group members. In a group with m individuals, each individual i assigns weight w_{ij} to individual j, with all $w_{ij} \geq 0$ and $\sum_{i=1}^{m} w_{ij} = 1$ for all i. This revision process is assumed to be iterative with a constant matrix of weights $\underline{W}$ and a vector of initial individual probability distributions, P, with elements $p_1, \ldots, p_m$. Then, after n iterations the vector of probabilities is

$$P^{(n)} = \underline{W}P^{(n-1)} = \underline{W}^n P.$$

The elements of $P^{(n)}$ will converge to the same limit; i.e., a consensus is reached, if and only if there is a vector $W^* = (w_1^*, \ldots, w_m^*)$ such that

$$\lim_{n \to \infty} w_{ij}^n = w_j^*$$

for all i and j, where w_{ij}^n is an element of $\underline{W}^n$. DeGroot proved that W^* exists if there is at least one person in the group who receives nonzero weights from all group members. The elements of W^* can be found

by solving the set of linear equations $w^* \underline{W} = w^*$ subject to the constraint

$$\sum_{j=1}^{m} w_j^* = 1.$$

The group probability distribution is then the linear combination of the initial individual assessments weighted by the w_j^*'s.

One specific advantage of the DeGroot formulation is that it explicitly reveals how weights are to be determined. Other justifications leave this question completely open. However, several procedures for assigning weights have been suggested and empirically tested, but will be discussed later since they pertain to other aggregation methods as well as the linear combination.

The linear combination is the only mathematical aggregation rule that has received much empirical attention as a means of generating composite probabilities. Several studies have shown that weighted linear combinations of individual probabilities are generally superior to individual assessments as evaluated by proper scoring rules (Brown, 1973; Stael von Holstein, 1971, 1972; Winkler, 1971). However, since proper scoring rules are concave functions on the probability simplex, the score of the average of individual probabilities will necessarily be better than the average of the individuals' scores. Nevertheless, the evidence is quite striking since usually only 10 percent or fewer of the individual subjects out-perform the group assessments.

Other evaluations also argue for the superiority of weighted linear combinations. Winkler (1971) made hypothetical bets based on both individual and weighted linear combinations of individual probability assessments for football game winners. For various betting schemes, bets based on the weighted linear combinations won from 2¢ to 47¢ more per

dollar bet than did bets based on individual assessments. This economic evaluation is rather impressive support for weighted linear combinations of individual probability assessments.

Bayesian models and approximations. Since probability assessments may be considered information pertaining to a set of hypotheses, a natural procedure for combining such assessments would be to use Bayes' Theorem, the formally correct procedure for combining probabilistic information. Somewhat similar treatments of this problem have been suggested by Dalkey (1975) and Morris (1974, 1977).

Morris derived results applicable from the point of view of a decision maker faced with the task of combining the probabilistic judgments of multiple experts with his or her own judgment. However, with some very minor adjustments, his model is applicable to the general problem of combining probabilistic judgments. Drawing on Bayes' theorem the most general form of the model is

$$p_G(\Theta) = k \cdot C(\Theta) \cdot p_1(\Theta) \cdot \cdots \cdot p_m(\Theta) \cdot p_0(\Theta)$$

where k is a normalization constant and p_0 is the prior probability, in most cases probably assumed to be uniform, but possibly derived from other sources; e.g., historical data and $p_i(\Theta)$ is the distribution assessed by expert i. $C(\Theta)$, the "Joint Calibration Function" (Morris, 1977), reflects both the lack of independence among the individual judgments in the sense that knowing the judgment of one individual provides information about the probable judgment of another individual; and the lack of calibration of the individual judgments. This function is generally impractical to derive because of the necessity for inordinate amounts of data or very complex judgments. Therefore, simplifying assumptions must be made to utilize this model.

The same problem occurs with Dalkey's (1975) development of the "probabilistic approach," which deals only with discrete probability assessments. In this model, the group probability of event Θ_k is derived as

$$p_G(\Theta_k) = \frac{\prod_{i=1}^{m} r_i(\Theta_k|p_i(\Theta_k))}{\sum_{j=1}^{n} D_{jk} \prod_{i=1}^{m} r_i(\Theta_j|p_i(\Theta_j))}.$$

Rather than aggregating $p_i(\Theta)$, the assessed probabilities, this formulation aggregates $r_i(\Theta_j|p_j(\Theta_j))$, the value of individual i's calibration function at $p_i(\Theta)$. For example, if for some assessor only 80 percent of the propositions assigned a probability of .9 occur, then r would be .8 when p is .9. The D_{jk} terms reflect the lack of independence of the individual judgments and the prior probabilities. These terms would often be very difficult to determine, and in many instances, the r_i's might also not be readily available.

Two major assumptions are necessary to make either of these models easily usable: independence among assessors and perfect calibration, i.e., $r_i=p_i$. Then, in the discrete case with uniform prior probabilities either model reduces to

$$p_G(\Theta_k) = \frac{\prod_{i=1}^{m} p_i(\Theta_k)}{\sum_{j=1}^{n} \prod_{i=1}^{m} p_i(\Theta_j)}. \tag{4}$$

If n, the number of hypotheses, is two, this model is equivalent to the likelihood ratio form of Bayes' Theorem, with each individual's odds as the likelihood ratio inputs. If the prior probabilities are not uniform, in Morris's model, the prior probability would simply be treated equivalently

to the assessment of another individual, but in the Dalkey model, the prior distribution enters into the calculations in a much more complex manner (see Dalkey, 1975, pp. 252-255).

With assessments of density functions, assumptions of independence and perfect calibration, and the additional requirement that all individually assessed densities be members of the same family of conjugate distributions; Morris' model becomes the natural-conjugate model suggested by Winkler (1968). Using conjugate distributions is not necessary for the Morris model, but does greatly simplify the mathematics.

Winkler generalized the conjugate model somewhat by allowing each individual's distribution to be weighted. Differences in weights should represent differences in the validity of the assessed distributions, while the sum of the weights (in this case not required to be one) should represent in some sense the independence of the assessments. Thus, Winkler argued for the sum of the weights to be between one and m, the number of assessors, because a sum of one represents complete dependence, while a sum of m represents complete independence. However, a type of dependence in which the entire set of distributions provides more information than do the single distributions by themselves might lead to sums greater than m, so such a restriction is really not justified.

The idea of weighting the individual assessments in the discrete case extends the model (eq. 4) to:

$$p_G(\Theta_k) = \frac{\prod_{i=1}^{m} p_i^{w_i}(\Theta_k)}{\sum_{j=1}^{n} \prod_{i=1}^{m} p_i^{w_i}(\Theta_j)} . \qquad (5)$$

where w_i is the weight assigned to individual i's assessment. If the weights are required to sum to one, the group probability is then the

normalized weighted geometric mean of the individual assessments. This model is then the multiplicative parallel to the linear combination model which is the weighted arithmetic mean. In considering the weighted geometric and arithmetic means, it is useful to keep in mind Dalkey's (1972) result showing that aggregation by addition generally destroys the multiplicative properties of the probabilities, whereas aggregation by multiplication destroys additive properties.

Pari-mutuel model. An ingenious and appealing aggregation model has been suggested by Eisenberg and Gale (1959) based on the pari-mutuel betting system used at race tracks. The pari-mutuel betting system provides a natural set of track or consensus odds (or equivalently, probabilities). Eisenberg and Gale investigated the conditions under which similar consensus probabilities could be explicitly determined from a set of individual assessments. They formulated the problem as follows. Suppose there are m individuals and n mutually exclusive and exhaustive events, and each individual i has amount b_i to bet, with the b_i's normalized to sum to one. Each individual i bets β_{ij} on event j,

$$\sum_{j=1}^{n} \beta_{ij} = b_i$$

so as to maximize his or her subjective expected value and the final consensus probabilities are proportional to the total amount bet on each event. That is

$$p_G(\Theta_k) = \sum_{i=1}^{m} \beta_{ik},$$

where equality holds because of the normalization of the b_i's. Individual i will maximize expected value by betting only on those events for which $p_i(\Theta_j)/p_G(\Theta_j)$ is maximum.

At this point the reasoning appears to be circular: individuals cannot bet without knowing p_G, and p_G cannot be determined until the bets are made. Eisenberg and Gale do not give a solution to this circularity. Rather, they simply prove that a set of bets and a unique set of consensus probabilities exist that are consistent with this model. The consensus probabilities are

$$p_G(\Theta_k) = \max_i \frac{b_i p_i(\Theta_k)}{\sum_{j=1}^{n} p_i(\Theta_j)\bar{x}_{ij}}.$$

The values $\bar{x}_{ij}$ are the values that maximize the function

$$F(x_{11},\ldots,x_{mn}) = \sum_{i=1}^{m} b_i \log \sum_{j=1}^{n} p_i(\Theta_j)x_{ij}$$

with $x_{ij} \geq 0$ and $\sum_{i=1}^{m} x_{ij} = 1$, for all i and j.

Norvig (1967) has proved the same result with a more intuitively appealing mathematical approach. He formulated the problem as an interactive process in which individuals place bets which lead to consensus probabilities, which then allow individuals to place new bets, etc. The consensus probabilities will then converge on the same probabilities specified in the Eisenberg-Gale model.

Weighting procedures. Most of the mathematical aggregation models allow the individual assessments to be differentially weighted. Even the pari-mutuel model, although not explicitly referring to weights, allows weighting via the amount each individual can bet. Therefore, the specification of weights is a necessary part of the use of these models. Several procedures have been suggested and empirically tested with linear

combination models, including both theoretically developed procedures and strictly *ad hoc* methods. In empirical tests, the theoretical procedures have not shown any superiority to *ad hoc* methods of assigning weights. An informal test (Hogarth, 1977) of weights derived using the DeGroot (1974) model showed it led to predictions that were slightly inferior to those of a simple average (equal weights).

Roberts (1965) has suggested another weighting procedure based on the predictive probability of previous assessments. However, because the weights for most individuals will rapidly approach zero, this procedure has proved to be impractical (Winkler, 1971).

The more *ad hoc* weighting procedures, usually based on past performance, self-ratings, or ratings by others, have received considerable attention. Stael von Holstein (1972) compared several weighting procedures based on prior performance and found little or no difference among them. Similar results have been obtained with self-ratings and ratings by others (Gough, 1975; Rowse, Gustafson, and Ludke, 1974; Stael von Holstein, 1971; Winkler, 1971).

These results are not surprising given the "flatness" of linear models (von Winterfeldt and Edwards, 1973). This flatness ensures that relatively large changes in weights will produce only small changes in the output of the model. Since both the aggregation procedure (linear combination) and the evaluation procedure (proper scoring rules) are linear models, flatness is doubly ensured. Whether or not this insensitivity to weights also holds for nonlinear aggregation models and other types of evaluations remains to be investigated.

Behavioral Interaction

An alternative to mathematical aggregation of individual

probabilities is some kind of behavioral interaction. This can be used either in conjunction with mathematical aggregation or simply by itself. Interaction here refers to any form of communication or transfer of information and ideas among the individuals making the assessments, so, therefore, is not limited to face-to-face discussions.

The most obvious reason for allowing interaction among group members is that each may have information that is useful to the others in making their assessments. By sharing this information the assessment of each individual, and, therefore, the group assessment may be improved. This need not necessarily happen, however, because the information may, in fact, produce worse assessments. However, if the potential can be exploited, the interaction should be beneficial. In fact, consensus probabilistic judgments determined through interaction have been shown empirically to be superior to individual judgments (Goodman, 1972; Stael von Holstein, 1971).

Social psychological research suggests some other reasons that favor interacting groups in a wide variety of judgmental tasks. Interaction is likely to make group members feel more responsible for the group judgment, and, therefore increase their motivation. This also has a practical beneficial side effect: the group members are more likely to accept a judgment arrived at in this manner as the basis for making a decision (Collins and Guetzkow, 1964; Davis, 1969).

Given these potential positive benefits of behavioral interaction, considerable interest has developed in finding ways to take advantage of them without the group being exposed to the known negative aspects of interaction such as dominant individuals and pressure for conformity that typically accompany uncontrolled interaction. In particular, two

procedures that control interaction have been developed and widely utilized in a variety of contexts: Delphi, developed by Dalkey and Helmer at The Rand Corporation; and the Nominal-Group-Technique (NGT) developed by Delbecq and Van de Ven at the University of Wisconsin. Both procedures rely on controlled interaction, and neither actually leads to a group consensus; therefore, necessitating the use of some type of mathematical aggregation. The procedural details and empirical support for these methods are reviewed in the following subsections.

Delphi. Delphi was first used in 1951 to elicit expert judgments about the number of A-bombs needed to reduce U.S. munitions output to a certain level (Dalkey and Helmer, 1963). Since then it has achieved wide-spread use, particularly in industry for predicting technological development (Linstone and Turoff, 1975; Sackman, 1974). Many different procedures have been used under the name "Delphi," but as originally conceived, Delphi includes three basic features: (1) anonymity of group members; (2) iteration of responses with controlled feedback between iterations; and (3) statistical aggregation (unspecified as to type) of individual judgments to form the group response (Dalkey, 1969b).

These characteristics are designed to reduce some of the potential problems associated with face-to-face discussion groups. The anonymity ensures that no individuals can dominate the group because of status. Iteration and controlled feedback allow the exchange of information without the value of the information being affected by its source. Finally, the statistical group response lessens the pressure for conformity and takes advantage of the error variance reduction of statistical aggregation.

The validity of responses obtained using Delphi was studied in

a series of experiments at The Rand Corporation (Dalkey, 1969a, 1969b; Dalkey, Brown and Cochran, 1970a, 1970b). In the only study that compared Delphi responses with the consensus of face-to-face discussion groups (Dalkey, 1969b), Delphi yielded more accurate answers on 13 of 20 questions, marginal support at best for Delphi. Additional support came from a second part of the study in which groups used Delphi between rounds one and two of responses, and face-to-face discussion between rounds two and three. There was slightly more improvement between rounds one and two, but again this support is quite weak given the small difference and the obvious design flaws. The Delphi procedure does lead to improved judgments with successive rounds, but the convergence of judgments is much larger. In fact, generally the judgments converge much more than is justified by the improvement (Dalkey, 1969a, 1969b).

The use of Delphi as a technique for generating quantitative assessments of unknown quantities from multiple experts seems to be much more extensive than can be justified by the empirical research (Sackman, 1974). Several features of Delphi can be questioned: the multiple iterations apparently produce more convergence than is justified; and the anonymity of respondents suppresses a potentially important feature of the feedback information; namely, its source.

Clearly, the value of Delphi has not been firmly established, particularly as a tool for assessing group probabilities. There have been enough positive results, however, to justify further investigations. A few studies have used the Delphi method to assess group probabilities. They will be discussed following the presentation of the Nominal-Group-Technique.

Nominal-Group-Technique. Van de Ven and Delbecq (1971) reviewed the literature on the effectiveness of nominal groups (groups with no spontaneous interaction) versus interacting groups on problem-solving and decision-making tasks, and concluded that a process combining the attributes of these two processes should be more effective than either alone. On this basis, they developed and tested the NGT. The specific procedure, described in Delbecq, Van de Ven, and Gustafson (1975), includes (1) silent judgments by individual group members in the presence of the group; (2) presentation to the group without discussion of all individual judgments; (3) group discussion for clarification and evaluation controlled by a group leader to prevent dominance and to focus on relevant issues; (4) individual reconsideration of judgments; and (5) mathematical aggregation of final individual judgments.

Thus, like the Delphi method, NGT may reduce pressure for conformity by not forcing a consensus. The controlled discussion also reduces the chance for dominance by individuals, although perhaps not to the extent Delphi's anonymity does. Both procedures eliminate the need for the group to provide structure since it is implicit in the procedure. The primary differences in Delphi and NGT are that NGT requires that group members actually be together physically and allows face-to-face discussion. NGT also provides knowledge of the source of any and all information. Additionally, NGT requires an active leader. Delbecq et al. (1975) discuss the advantages and disadvantages of this type of leadership role.

Much of the empirical support for the NGT comes from a problem-solving study with rather weak evaluation criteria (Van de Ven and Delbecq, 1974). Groups using Delphi, NGT, and uncontrolled interaction

were compared on the number of alternatives generated and the perceived satisfaction of group members. NGT clearly led to more satisfaction, while NGT and Delphi groups both produced more alternatives than the interacting groups. Neither of these measures has much relevance to the quality of the group judgments, but the satisfaction may be practically important.

Experimental comparisons with probabilistic judgments. Although neither Delphi nor NGT were developed for assessing probabilities, both obviously could be applied in this capacity. In fact, three studies have specifically compared these procedures with interacting groups and mathematical aggregation without any interaction. Gustafson, Shukla, Delbecq, and Walster (1973) compared groups making judgments about the likelihood ratios of male versus female given certain heights. Four types of groups were used: mathematical aggregation without interaction; NGT; Delphi; and modified interacting groups. The modification to the interacting groups was that no actual consensus was required prior to individual judgments after the interaction. Thus, interacting groups differed from NGT groups only in that NGT groups made individual judgments before the interaction. Geometric means were used to aggregate the individual likelihood ratio judgments. Using the average deviations of the group judgments from the true likelihood ratios, NGT groups produced the best assessments and Delphi groups, the worst.

A study by Gough (1975) used the same four types of groups as used by Gustafson et al. with the exception that the interacting groups made individual assessments prior to interaction and actually had to reach a consensus during the interaction. The assessments were five fractiles of the individuals' cumulative subjective probability

distribution for general information questions and a linear aggregation model was used. A quadratic proper scoring rule was applied to evaluate the probability distributions. Although Gough's results indicated that NGT groups produced the best assessments, the differences were quite small and probably did not justify his conclusions favoring the NGT.

The third study (Fischer, 1975) used the same types of groups as Gough with a different type of assessment. Subjects were asked to assess the probability of freshmen GPA's falling into four mutually exclusive and exhaustive categories given information about gender, high school GPA, SAT math, and SAT verbal scores. Fischer's evaluation method, a logarithmic proper scoring rule was similar to Gough's, as were his results. There was virtually no difference among the groups. Fischer attributes much of the difference between his results and those of Gustafson et al. to the dependent variable used to evaluate the assessments. His basic argument is that large differences in likelihood ratios may be only small differences when transformed into probabilities, particularly at the extreme ends of the probability scale. Thus, it appears results may very much depend on the way in which they are evaluated.

AN EXPERIMENTAL COMPARISON AND EVALUATION

As suggested in the Introduction, several questions about how group probabilities should be assessed need to be answered empirically. The previous section outlining the mathematical and behavioral interaction approaches to assessing group probabilities and related literature indicates that these questions have yet to be answered. This experiment attempts to answer these questions.

The first question is whether interaction of some kind will improve the group probabilities compared with probabilities derived by mathematically aggregating the individual assessments. If interaction does improve assessments, what type of interaction allows the most improvement? In this study four types of interaction were used, along with a no interaction condition. They represent the interaction processes typically found in previous research: Delphi, NGT, and interacting groups forced to reach a consensus (hereafter called consensus or CON groups), along with a fourth process (MIX) that is somewhat a mixture of Delphi and NGT. This process, like NGT, has individuals make judgments and present them to the group, but allows only presentation of specific reasons for the judgment without open discussion. In this respect, it is more similar to Delphi. These interaction processes represent a continuum with respect to the latitude the groups have for interacting ranging from none to complete freedom.

Another issue investigated is the differences in group probabilities caused by use of different mathematical aggregation models. Because different models can be used depending on whether the assessed probabilities are discrete or continuous, both types of assessments were obtained. The basic aggregation models that were used included the linear combination model for both discrete and continuous probabilities, the conjugate model with weights summing to one and to m (the number of group members), for continuous probabilities, the discrete counterparts of the conjugate model--weighted normalized geometric mean and aggregation by likelihood ratios--and the pari-mutuel model for discrete probabilities. Additionally, three sets of weights were used with each model that allows for differential weighting: weights obtained from the DeGroot (1974) model; weights reflecting each individual's self-rating relative to the self-ratings of other individuals; and equal weights. Group probabilities derived by aggregating individual probabilities with these models can also be compared with the consensus probabilities decided upon by CON groups.

An additional product of this study is a comparison of individual and group probability assessments using primarily extremeness, calibration, and proper scoring rules. A quadratic scoring rule was used for discrete assessments. Continuous assessments were evaluated with a linear transformation of the continuous ranked probability score,

$$S^* = (S_u - S)/S_u,$$

where S is the usual score (eq. 2) and S_u is the score for a uniform distribution; i.e. $p(\Theta) = \Theta$. This permissible transformation makes the scores easier to interpret since S* does not depend on the

true value as S does. The range of S^* is from -4.0 to 1.0 with a uniform distribution receiving a score of 0.0.

Experimental Method

Subjects. Eleven four-person groups were used. Ten groups participated in the assessment of discrete probabilities, but one of these groups was unavailable to assess continuous probabilities so was therefore replaced. This causes no problem in data analyses since the data from the two types of assessments are analyzed separately. The subjects were predominantly graduate students at the University of Southern California or their friends. Each subject was familiar with the other three members of the group. Subjects were paid $20 plus bonuses based on evaluations of some of their responses with the proper scoring rules for each of the two sessions, bringing total payment to approximately $5 to $6 per hour.

Stimuli. For the discrete assessments, the stimuli were 100 two-alternative general information questions randomly sampled from a collection of about 700 such questions.[1] These questions were randomly divided into five sets of 20 questions.

The continuous stimuli were general information questions about percentages. A set of these questions was developed with five true values falling into each range of 5 percent from 10 percent to 40 percent and from 60 percent to 90 percent and two true values in each 5 percent range between 40 percent and 60 percent. These questions were randomly assigned to five sets of questions so that each set had one true value in each 5 percent range from 10 percent to 40 percent and from 60 percent

[1] I would like to thank Sarah Lichtenstein for making these questions available.

to 90 percent and two true values in each 5 percent range between 40 percent and 60 percent. This was to ensure that differences in the quality of probabilities assessed for the different question sets were not due to the true values of the questions. The very extreme percentages were avoided because of the large biases usually found in assessed distributions for these questions (Fujii, Seaver, and Edwards, 1977).

Procedure. Each group of subjects participated in two sessions: the first assessing discrete probabilities and the second, continuous probabilities. Sessions lasted from three to four and a half hours with the continuous assessment session taking about an hour longer than the first session because additional training was needed. Each group answered a different set of questions in each of the five interaction conditions. The question sets and the order of interaction conditions were balanced in a 5 x 5 Greco-Latin square.

For the discrete assessments, subjects were required to choose the answer they thought was most likely to be correct and then assess the probability ($p \geq .5$) that the choice was correct. Also, for each question they were instructed to assign weights to each group member reflecting their belief about how much each group member's opinion should contribute to the "group opinion." These weights were to reflect subjects prior beliefs about the expertise of the group members with respect to the question under consideration. Each individual whose opinion should contribute nothing was assigned a weight of zero. Of the remaining group members, a weight of 10 should be assigned to those whose opinion should contribute the least. Any remaining individuals should be assigned weights reflecting their contribution relative to those receiving weights of 10. For example, if another individual's opinion should contribute five times as much, that individual would receive a weight of 50. Weights were assigned for each question

during both the initial and final probability assessments in all interaction conditions.

Subjects were then given a sheet of paper showing the quadratic scoring rule that would be used to evaluate their assessments. In addition to a fixed payment of $20, subjects could win or loose money based on applying the scoring rule to judgments on two randomly selected questions from each set of 20. The paper included the amount to be won or lost for probabilities between .5 and 1.0 in steps of .05 plus .99. The quadratic scoring rule (eq. 1) was linearly transformed so that any assessment of .5 would mean nothing won or lost, while an assessment of 1.0 would result in a win of one dollar if the choice was correct, or a loss of three dollars if the choice was wrong. Four sample questions were answered by each subject and the answers to these questions were scored to illustrate the scoring rule.

The procedure for the initial individual assessments was the same for each interaction condition. The subjects answered each of the 20 questions without any discussion among themselves. After all group members had completed these questions, the procedure varied depending on the interaction condition. Table 1 shows the major differences in the interaction conditions.

TABLE 1

MAJOR DIFFERENCES IN TYPES OF INTERACTION

Type of Interaction	Reconsider with Information about Other Judgments	Knowledge of Judgment Source	Verbal Information Exchange	Uncontrolled Discussion	Consensus Necessary
None					
Delphi	Yes				
MIX	Yes	Yes	Yes		
NGT	Yes	Yes	Yes	Yes	
CON	Yes	Yes	Yes	Yes	Yes

In the no interaction condition, the subjects were simply told the answers and scored two pre-selected questions.

In the Delphi condition, the experimenter collected the assessments and explained that the subjects would have two subsequent chances to reassess their probabilities, each time with additional information about the assessments of the other group members. For each question the subjects were given the four assessments of the group without any information about who made which assessment. On the basis of this information they reconsidered their judgments. In addition, they were instructed to write any information that might be useful to other group members in space provided on the answer sheets. In particular, if someone's judgments differed radically from other group members, that person should attempt to explain the reasoning behind the judgment. After all 20 questions were again answered, the same process was repeated with the feedback, including any written information provided by the subjects. After the final set of responses was completed, the answers were given and two questions were scored from each of the initial and final assessments.

In the MIX condition, each group member presented his or her assessment for the question under consideration to the group verbally. After each assessment had been presented, any group member was allowed to state any reasons underlying the assessment or any information that might be useful to other group members. Each individual then reconsidered the assessment for that question. After all questions had been considered a second time, the answers to all questions were given and two assessments from each of the initial and subsequent assessments were scored for pay.

The NGT groups were the same as the MIX groups except after presentation of the individual assessments, a general face-to-face discussion was allowed with only the restriction that it be relevant to the question under consideration.

The CON groups differed from the NGT groups in that the presentation of individual judgments was not required and the groups had to reach consensus (agreement) about the assessment.

The second session, in which continuous probabilities were assessed, was similar in most respects to the first except considerably more training for the assessments was provided. For these questions subjects were requested to assess two parameters of a beta distribution representing their opinion about the possible answers to questions involving percentages. Rather than asking for α and β, the usual beta parameters, the parameters of $m = (\alpha - 1)/(\alpha + \beta - 2)$, the mode, and $n = \alpha + \beta - 2$, which reflects the tightness of the distribution and can be considered as a sample size, were assessed. To teach the subjects the correspondence between these parameters and the actual shape of the probability distributions, each subject was given a book containing graphs of the density and cumulative distribution functions of beta distributions with values of m beginning at .05 and increasing by steps of .05 to .95, and values of n equal to 0, 2, 5, 10, 15, 20, 25, 30, 50, 75, and 100 for each m. Each of the graphs also included the corresponding numerical quantities of density and cumulative probability for each .05 increment. Subjects kept these books for reference throughout the session.

After the meaning of the graphs and the correspondence between the parameters and the shape of the distributions was explained, a test was made to ensure that the subjects knew this correspondence. Subjects

were presented graphs of various beta densities and cumulative distributions together and asked to estimate the parameters of the distributions. Actually, only the parameter n was estimated since subjects had no trouble with the correspondence between m and the distributions. These graphs were presented to the subjects individually until 12 consecutive estimates (each subject three times) of n were between 2/3 and 3/2 of the true value. The total number of graphs presented ranged from 89 to 208 for the various groups.

After the training, subjects were instructed about the scoring rule to be used for these assessments, given four practice questions and answers, reminded of the procedure for assessing weights, and began the task with the interaction conditions. Following the completion of the second session, subjects were questioned as to which procedure they would prefer to use if they were in a real decision making group which needed to determine some relevant probability.

Results

Discrete assessments. The average quadratic scores of various aggregation models both before any interaction and after each of the types of interaction are presented in Table 2(a), along with the average individual scores and the average score of the actual consensus assessments. The aggregation models are the linear model (eq. 3), the geometric mean model (eq. 5), and the likelihood ratio model (eq. 4). The three weighting procedures are equal, DeGroot (1974), and self-rating, derived by first normalizing the weights assigned by each individual to sum to one and then again normalizing the (normalized) weights individuals assigned to themselves.

$$w_i = \frac{\dfrac{w_{ii}}{\sum_j w_{ij}}}{\sum_i \left[\dfrac{w_{ii}}{w_{ij}}\right]}$$

where w_i is the derived self-weight for individual i and w_{ij} is the weight assigned by individual i to individual j.

The most notable result in Table 2(a) is that the likelihood ratio model does quite poorly. The linear and gemoetric mean models differ only slightly as do the weighting procedures. Also, interaction does not generally seem to have much effect on the group scores, although the NGT scores tend to be somewhat higher, but it increases the individual scores. An analysis of variance with five repeated measures factors: interaction type, questions, repetition (before or after interaction), aggregation model (only linear and geometric mean), and weights, generally confirmed these conclusions. Other than the questions factor, which is of little interest here, no main effects were significant and only two interaction terms were significant: the aggregation model by weights interaction, $F(2,18)=11.7$, $p \leq .001$, and the repetition by aggregation model by weights interaction, $F(2,18)=8.48$, $p \leq .003$.

Although the evaluation with the scoring rule shows little difference among the group probabilities, other characteristics show more distinct effects. Table 2(b) shows the average probabilities assigned to the correct response, a measure of the extremeness of the assessments. The group probabilities are more extreme than the individual probabilities, the likelihood ratio model produces the most extreme probabilities and interaction leads to more extreme probabilities. An analysis of variance confirmed the effects apparent in the means showing the probabilities to be more extreme after interaction, $F(1,9)=30.5$, $p \leq .001$, more extreme with the geometric mean than the linear model, $F(1,9)=29.2$,

TABLE 2(a)

AVERAGE QUADRATIC SCORES

	Linear			Geometric Mean			Likeli-hood Ratio	Actual Consen-sus	Indi-vidual
	Equal	Self-rating	De-Groot	Equal	Self-rating	De-Groot			
Before	.562	.565	.572	.570	.551	.568	.495		.494
After	.577	.569	.573	.557	.545	.549	.449	.556	.541
Delphi	.573	.557	.577	.541	.529	.542	.447		.529
MIX	.565	.558	.554	.547	.528	.527	.429		.526
NGT	.599	.595	.599	.584	.576	.582	.465		.556
CON	.572	.564	.562	.555	.547	.544	.454		.551

TABLE 2(b)

AVERAGE PROBABILITY ASSIGNED TO CORRECT ANSWER

	Linear			Geometric Mean			Likeli-hood Ratio	Actual Consen-sus	Indi-vidual
	Equal	Self-rating	De-Groot	Equal	Self-rating	De-Groot			
Before	.552	.575	.568	.590	.604	.601	.636		.552
After	.613	.623	.620	.631	.634	.633	.655	.635	.613
Delphi	.605	.611	.614	.621	.621	.626	.655		.605
MIX	.594	.607	.602	.617	.620	.616	.638		.594
NGT	.627	.639	.638	.648	.653	.654	.667		.627
CON	.626	.633	.627	.637	.641	.635	.660		.626

$p \leq .001$, and less extreme with equal weights, $F(2,18)=12.5$, $p \leq .001$. In addition, the three two-way interactions among these three factors were significant. However, neither the main effect due to interaction type, nor any of the interactions with that factor were significant.

Calibration, another desirable feature of probabilities, also showed some differences. Figure 1 shows the calibration curves for group and individual probabilities, both before and after interaction. The group probabilities are aggregated over both the linear and geometric mean models, all three weighting procedures, and all interaction types. Group probabilities are clearly better calibrated than individual probabilities before interaction, but interaction causes the calibration of the group probabilities to get worse while improving the calibration of the individual probabilities.

Neither weighting procedures nor type of interaction had any notable effect on calibration, so the calibration curves for the aggregation models shown in Figure 2 before interaction, and Figure 3 after interaction are aggregated over those variables. The linear model leads to quite well-calibrated probabilities before interaction, while the likelihood ratio model produces very poor calibration.

The use of the pari-mutuel model for aggregating individual probabilities had to be limited for cost reasons. To aggregate the probabilities of all groups for all questions using all weighting procedures would have required over 100 hours of cpu time. To reduce this computation to a more realistic level, one of the ten groups was randomly selected and the pari-mutuel model was used to aggregate the individual assessments of that group. Table 3 gives the mean quadratic scores and mean probabilities assigned to the correct response for the assessments of this group only. The pari-mutuel model generally produced lower scores and less extreme probabilities than the linear or geometric mean models. The

Figure 1

INDIVIDUAL VERSUS GROUP CALIBRATION: DISCRETE ASSESSMENTS

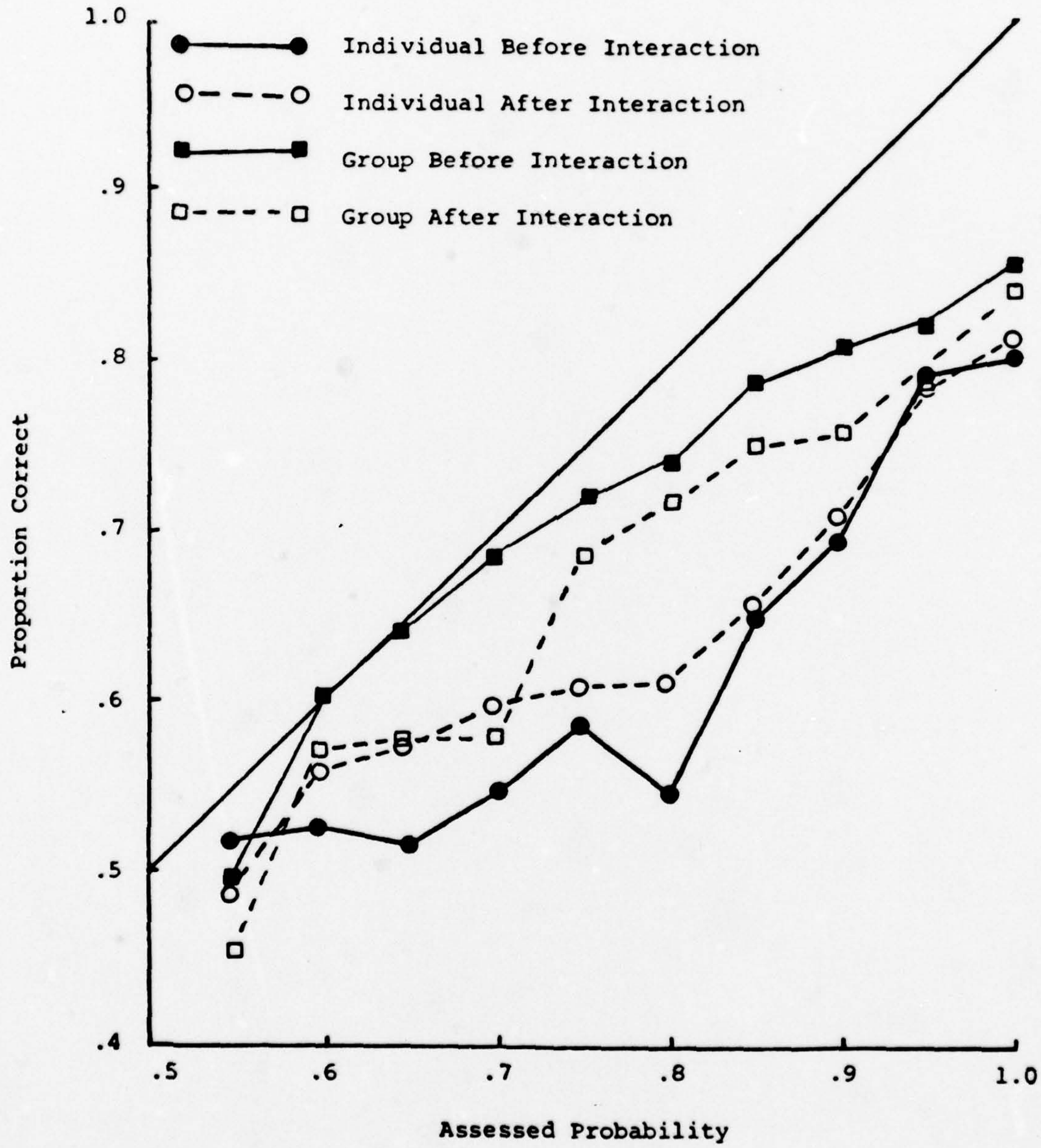


Figure 2

CALIBRATION OF DIFFERENT AGGREGATION MODELS BEFORE INTERACTION: DISCRETE ASSESSMENTS

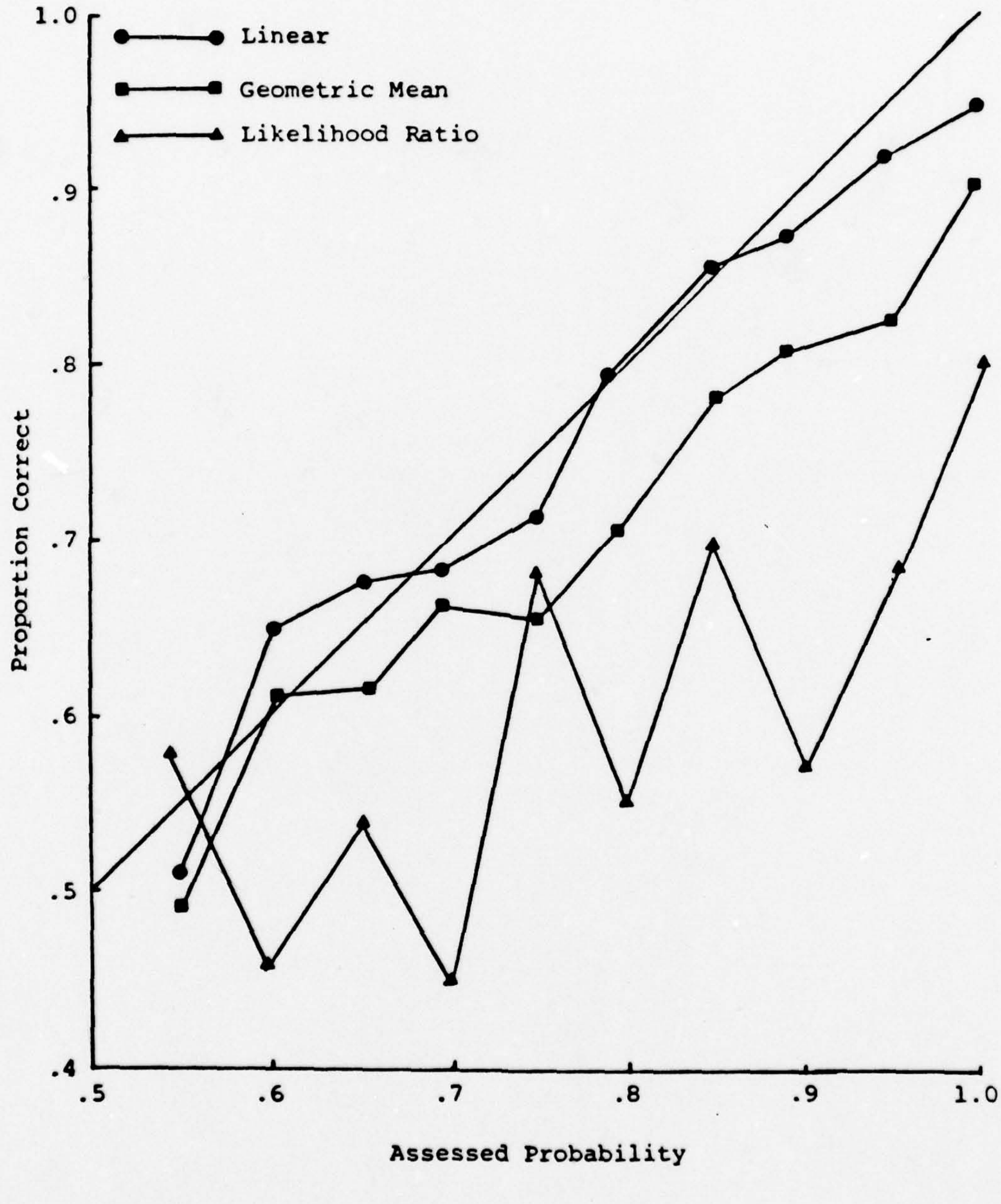


Figure 3

CALIBRATION OF DIFFERENT AGGREGATION MODELS AFTER INTERACTION: DISCRETE ASSESSMENTS

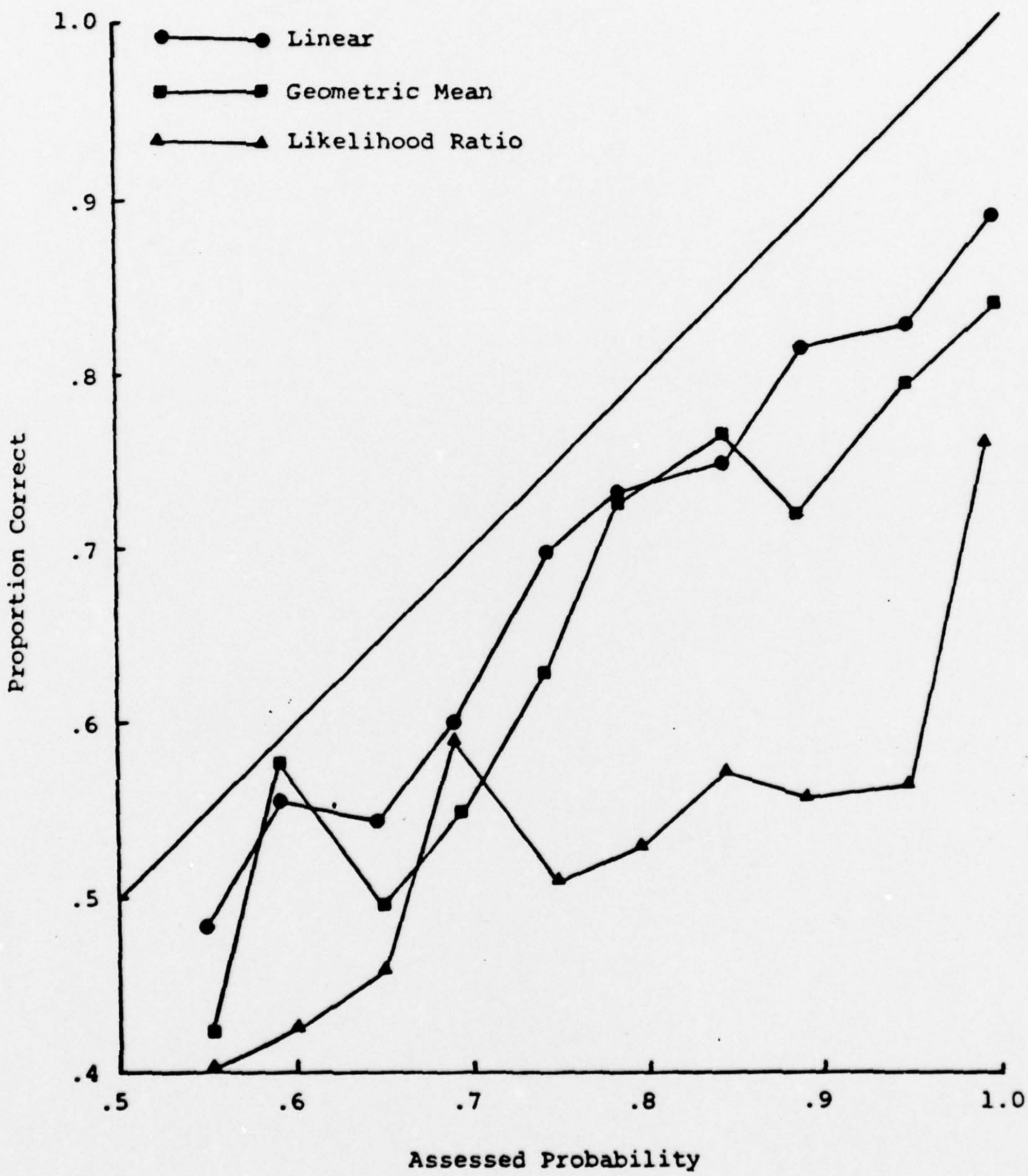


TABLE 3(a)

AVERAGE QUADRATIC SCORES FOR SINGLE RANDOMLY SELECTED GROUP

Before Interaction				After Interaction			
	Weights				Weights		
Model	Equal	Self-rating	DeGroot	Model	Equal	Self-rating	DeGroot
Linear	.612	.643	.618	Linear	.658	.668	.663
Geometric Mean	.639	.638	.636	Geometric Mean	.673	.676	.674
Pari-Mutuel	.546	.596	.556	Pari-Mutuel	.605	.630	.608

TABLE 3(b)

AVERAGE PROBABILITY ASSIGNED TO CORRECT ANSWER FOR SINGLE RANDOMLY SELECTED GROUP

Before Interaction				After Interaction			
	Weights				Weights		
Model	Equal	Self-rating	DeGroot	Model	Equal	Self-rating	DeGroot
Linear	.578	.622	.586	Linear	.634	.650	.643
Geometric Mean	.626	.654	.631	Geometric Mean	.656	.666	.661
Pari-Mutuel	.532	.576	.539	Pari-Mutuel	.588	.611	.594

relatively higher scores and more extreme probabilities using self-rating weights are apparently an anomaly of this particular group.

Figure 4 shows the pari-mutuel calibration curves for this group, along with the calibration of the linear model for reference curves. Given the irregularity of the curves and the small samples on which they are based (each point represents about 40 to 60 assessments), the calibration resulting from the use of the pari-mutuel model does not appear to be systematically different from the linear model.

Since assessments tended to become more extreme after interaction, some of the factors that might affect changes in probability assessments were examined. Four types of qualitative changes were considered: switches to the other answer; less extreme assessments; no change; and more extreme assessments. The factors considered were the split of initial individual answers, all agree (4-0), 2-2, and 3-1 for both the three individuals and the single individual; the type of interaction; the individual's probability relative to those given by group members selecting the other answer; and the individual's probability relative to those given by group members selecting the same answer. The latter two factors were divided into three categories, larger than all the other probabilities, between or equal to the other probabilities, or smaller than all the other probabilities. Table 4 presents the conditional percentages of changes for the given levels of each of these factors.

Changes generally display the intuitively expected patterns. The more other group members agree with an individual, the less likely that individual is to switch answers, and the judgment is more likely to become more extreme. Switches are more likely for individuals with probabilities smaller than those both with whom they agree and with whom they

Figure 4

CALIBRATION OF SINGLE RANDOMLY SELECTED GROUP INCLUDING PARI-MUTUEL MODEL: DISCRETE ASSESSMENTS

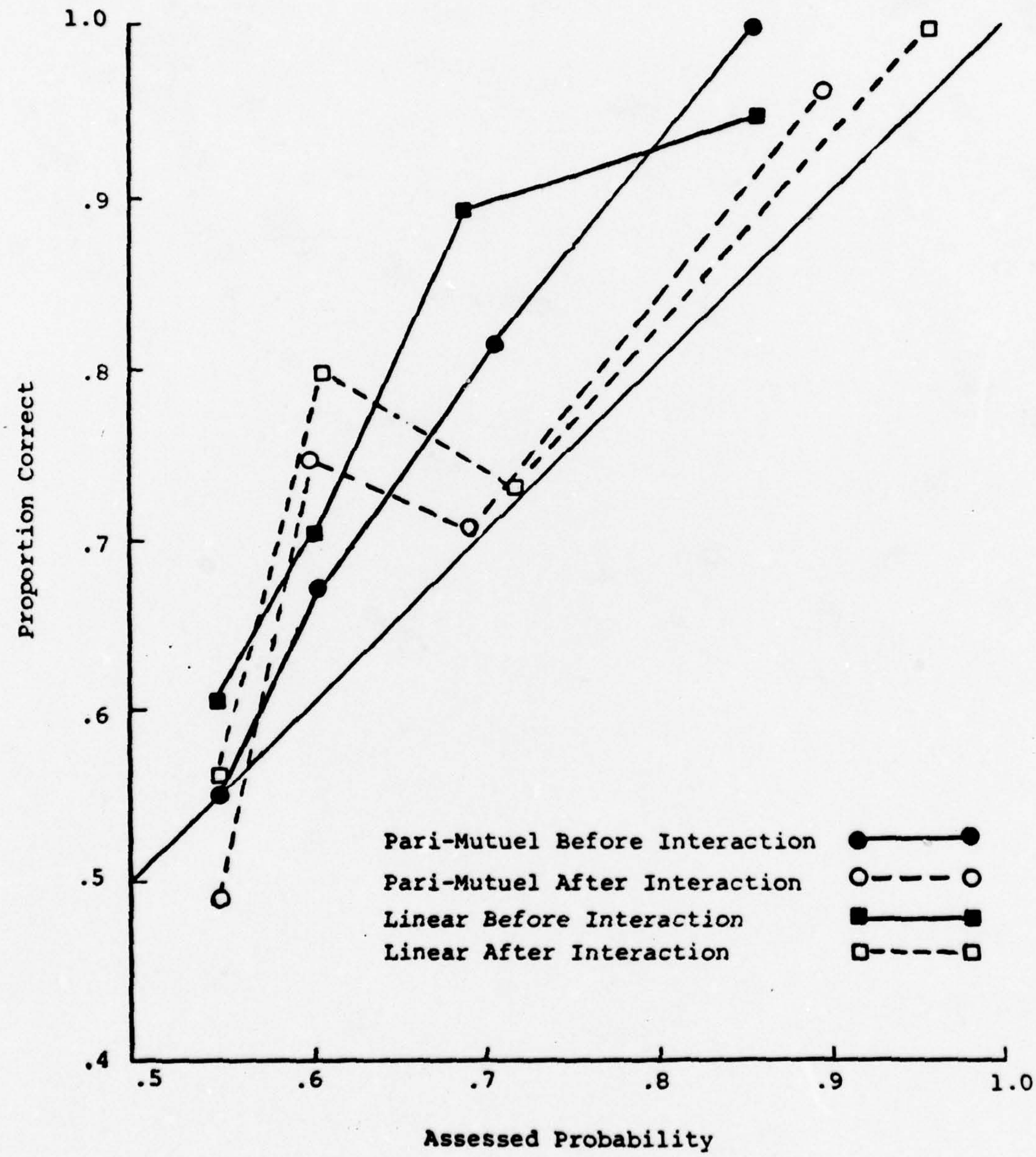


TABLE 4

CONTINGENCY TABLES FOR CHANGES IN INDIVIDUAL JUDGMENTS
(Percentages)

Split of Initial Individual Judgments	Change: Switch	Change: Less Extreme	Change: Same	Change: More Extreme	N
4 - 0	1.2	4.8	35.9	58.0	808
3 - 1	15.3	8.9	34.4	41.4	1134
1 - 3	47.9	12.4	27.5	14.0	378
2 - 2	31.9	13.2	33.1	21.8	880
Marginal Mean	20.2	9.5	33.8	36.5	3200
Type of Interaction					
Delphi	20.4	11.9	25.8	41.9	800
MIX	17.3	7.6	41.8	33.3	800
NGT	20.1	7.8	37.0	35.1	800
CON	23.0	17.1	29.0	30.9	800
Marginal Means	20.2	9.5	33.8	36.5	3200
Compared to Probabilities for Other Answer					
Larger	10.1	15.8	45.7	28.4	810
Between or Equal	35.0	11.5	28.3	26.2	820
Smaller	35.0	5.6	23.1	36.3	762
Marginal Means	26.6	11.0	32.5	29.9	2392
Compared to Probabilities for Same Answer					
Larger	17.0	28.1	45.5	9.4	814
Between or Equal	12.3	5.2	36.2	46.3	1237
Smaller	22.7	3.9	19.8	53.6	771
Marginal Means	16.5	9.1	34.4	40.0	2822

disagree. Among individuals who agree, there is a tendency toward averaging with the largest assessments remaining the same or getting less extreme, while the middle assessments remain the same or become more extreme, and the smallest assessments become more extreme.

These aggregated tables mask some of the more striking effects. For example, with a 4-0 split, 92 percent of the subjects with the smallest assessments became more extreme. Or with a 3-1 split, 68 percent of the single individuals switched if their probability was smaller than those of the individual with whom they disagreed, while only 20 percent switched if their probability was larger. These tables also conceal a non-intuitive interaction: among the three agreeing individuals in a 3-1 split, individuals with assessments less than or equal to the assessment of the individual who disagreed were more likely to switch if their assessment was larger than the assessments of the agreeing individuals (32%) than if it was equal to or between (22%) or smaller (21%).

Overall, interaction did produce a convergence in judgments. The standard deviations of individual judgments were reduced by an average of 25 percent, 26 percent, 27 percent, and 53 percent after Delphi, MIX, NGT, and CON interactions respectively.

Continuous assessments. Table 5(a) shows the mean scores of both individual and group assessments with the various aggregation models, weighting procedures, and types of interaction. The two aggregation models are the linear model and the conjugate model with weights summing to one. The linear model group probabilities were calculated by averaging the distributions at each step of 5 percent. All scores were computed by assuming a linear cumulative distribution between each 5 percent step. These approximations were necessary to reduce computation time.

TABLE 5(a)

AVERAGE SCORES FOR CONTINUOUS ASSESSMENTS

	Linear			Conjugate			Actual Consensus	Individual
	Equal	Self-rating	DeGroot	Equal	Self-rating	DeGroot		
Before	.005	-.011	.003	-.058	-.080	-.064		-.186
After	-.063	-.077	-.072	-.112	-.120	-.119	-.016	-.035
Delphi	-.036	-.057	-.036	-.075	-.099	-.079		-.011
MIX	-.094	-.100	-.113	-.160	-.147	-.172		-.114
NGT	-.012	-.026	-.016	-.066	-.078	-.072		.006
CON	-.112	-.125	-.123	-.146	-.158	-.155		-.022

TABLE 5(b)

AVERAGE DENSITY FOR CORRECT ANSWER

	Linear			Conjugate			Actual Consensus	Individual
	Equal	Self-rating	DeGroot	Equal	Self-rating	DeGroot		
Before	1.71	1.72	1.74	2.05	2.06	2.07		1.73
After	2.03	1.99	1.99	2.08	2.07	2.09	2.20	2.00
Delphi	2.20	1.98	1.98	2.15	2.04	2.14		2.01
MIX	1.84	1.87	1.83	1.85	1.93	1.84		1.85
NGT	1.97	1.96	2.00	2.09	2.10	2.12		1.99
CON	2.13	2.13	2.15	2.23	2.23	2.24		2.16

TABLE 5(c)

AVERAGE IQ RANGE FOR CONTINUOUS ASSESSMENTS

	Linear			Conjugate			Actual Consensus	Individual
	Equal	Self-rating	DeGroot	Equal	Self-rating	DeGroot		
Before	.266	.254	.254	.126	.124	.124		.140
After	.157	.152	.152	.113	.112	.113	.096	.115
Delphi	.161	.155	.155	.113	.113	.113		.115
MIX	.182	.175	.176	.118	.116	.119		.118
NGT	.162	.157	.156	.117	.114	.115		.120
CON	.121	.120	.119	.104	.104	.105		.108

Subjects did not generally do well on this task as shown by the scores being negative; i.e., worse than the score obtained with a uniform distribution. Correlations between the individual assessed modes and the true values were only .37 before interaction and .50 after interaction, at best a moderate relationship. In addition, the assessed values of n and the error measured by the absolute difference between the mode and the true value were not related ($r=-.06$ before interaction and $-.03$ after interaction).

Interaction lowered the scores of the group probabilities $F(1,9)=6.19$, $p \leq .035$, while raising those of individuals. In fact, after interaction, the individuals had higher scores than the groups. The best scores were received by the actual consensus judgments. However, inspection of the means indicates the differences are rather trivial in size. The repetition by model interaction was also significant, $F(1,9)=32.3$, $p \leq .001$, but again the differences were rather small. As for discrete assessments, neither the type of interaction nor the weights used made a difference in the scores.

Table 5(b) shows the mean densities at the true values. Again, the assessments became more extreme after interaction, $F(1,9)=13.3$, $p \leq .005$, and aggregation with the conjugate model leads to higher densities, $F(1.9)=130.0$, $p \leq .001$. The interaction between these factors was also significant, $F(1,9)=34.0$, $p \leq .001$.

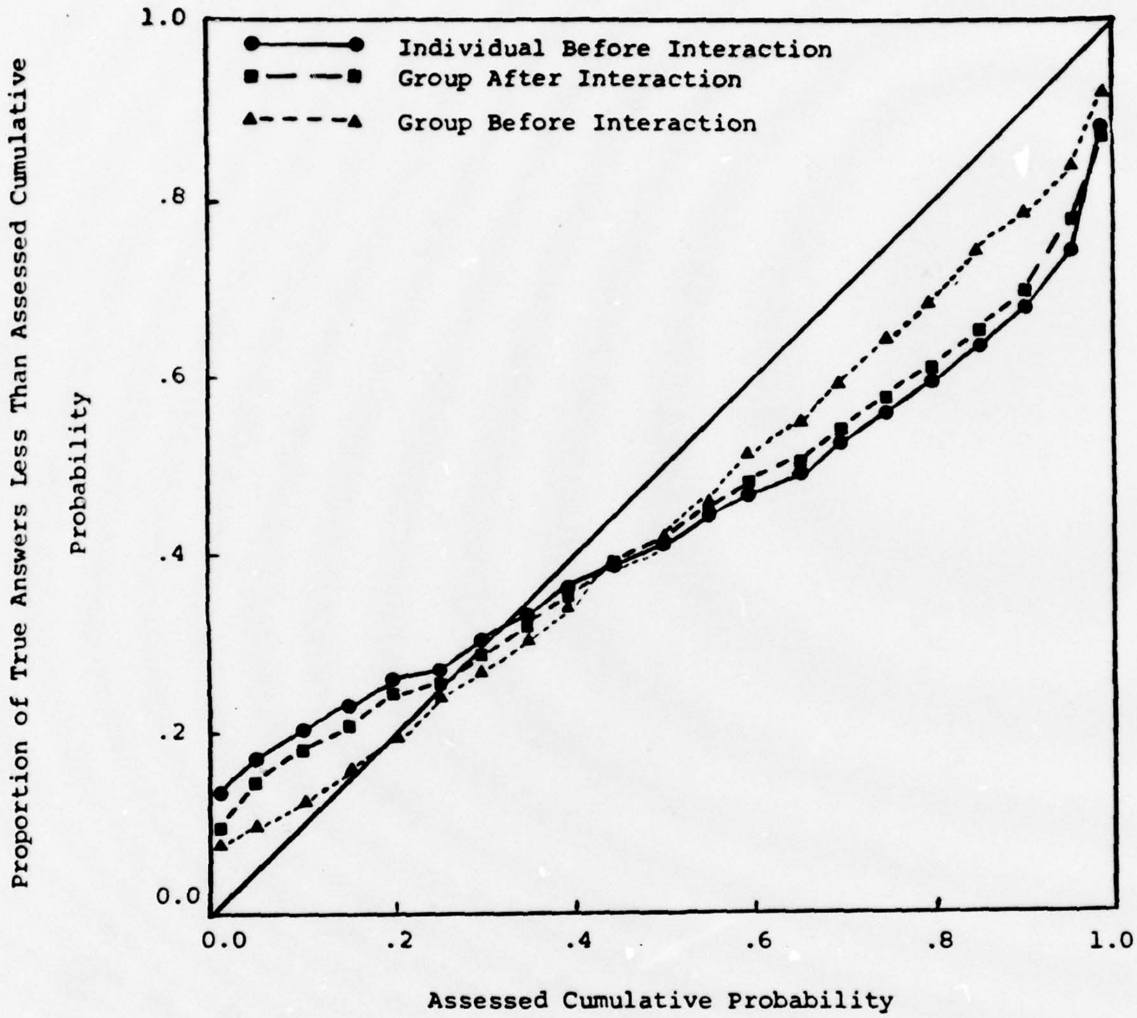
Another characteristic of the continuous probability assessments that reflects their extremeness, but not necessarily their accuracy, is their dispersion. Table 5(c) shows the mean values of one measure of dispersion, the interquartile (IQ) range. Group distributions derived

with the linear model ovbiously have larger dispersion than those from the conjugate model, $F(1,9)=161.2$, $p \leq .001$. And interaction considerably reduces the IQ range, $F(1,9)=164.2$, $p \leq .001$, particularly with the linear model (repetition x aggregation model interaction, $F(1,9)=130.2$, $p \leq .001$). Additionally, the CON interaction produced smaller IQ ranges (type of interaction main effect, $F(3,27)=5.51$, $p \leq .004$), indicating generally more agreement as a result of this type of interaction. Also here, surprisingly, the weights made a significant, $F(2,18)=8.73$, $p \leq .002$, although not substantial difference: equal weights produced more dispersed distributions. All the two-way interactions among repetitions, aggregation model, and weights were significant, although all except the repetition by aggregation model were relatively less substantial than the main effects.

How well calibrated are the continuous assessments? Figure 5 shows the calibration curves for the individual assessments before interaction and the group assessments both before and after interaction aggregated across weighting procedures, linear and conjugate models, and all types of interaction. Individual calibration after interaction is not shown because it differs little from the calibration before interaction (maximum vertical difference in curves = .013). These curves plot the percentage of true values (ordinate) falling below the specified value of the cumulative distribution (abscissa). Perfect calibration would result in a straight line from (0,0) to (1,1). The specific percentages of true values falling in the tails (less than .01 or greater than .99) and in the IQ range of the distributions are tabulated in the figure. These values are often used to measure the calibration of continuous assessments when the entire distributions are not assessed. All the distributions tend to be too tight with

Figure 5

INDIVIDUAL VERSUS GROUP CALIBRATION: CONTINUOUS ASSESSMENTS



	Tails	IQ Range
Individual Before Interaction	25.1%	28.7%
Individual After Interaction	24.9%	27.6%
Group Before Interaction	15.1%	40.3%
Group After Interaction	22.1%	31.7%

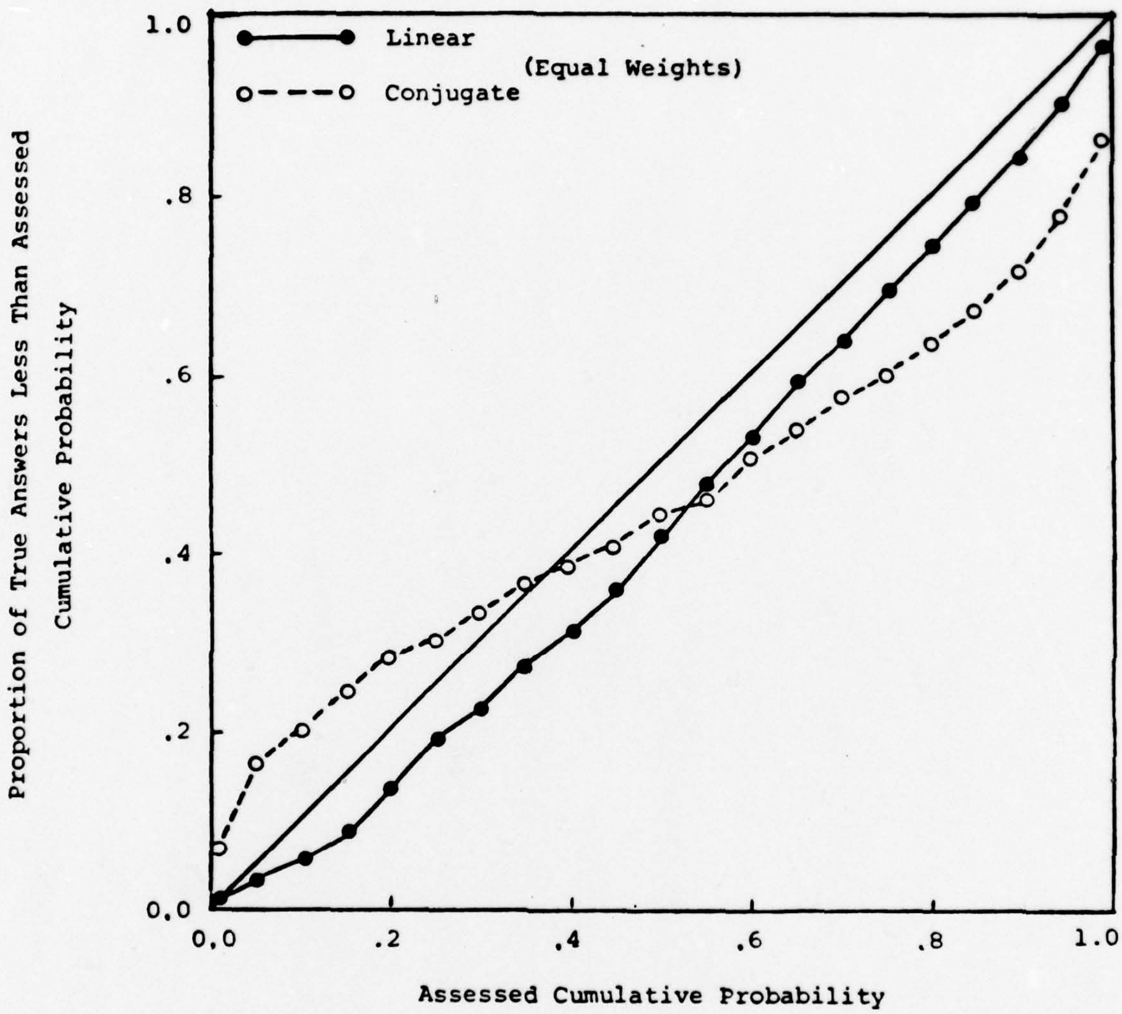
too many true values in the tails and too few in the IQ ranges. The group distributions, however, are better calibrated than the individual distributions, although as with the discrete probabilities, interaction leads to poorer calibration for the group distributions. Also, interestingly, the curves are not symmetric: the assessed distributions are displaced to the left of the true value more often than to the right.

The type of interaction again had little effect on calibration: percentages of true value in the tails ranged from 18 percent for NGT groups to 27 percent for CON groups; and IQ range percentages ranged from 29 percent for CON groups to 35 percent for Delphi. But as shown in Figures 6 and 7, the aggregation model did affect calibration both before and after interaction. The calibration curves are plotted only for equal weights since the curves for other weighting procedures are very similar (see Figures for maximum discrepancies). The group probabilities derived with the linear model are clearly better calibrated than those from the conjugate model. In fact, before interaction the linear model probabilities are very well calibrated, except for a slight underestimation displacement. Otherwise, the group distributions are too tight (too many true values in the tails and too few in the IQ range) and all are generally displaced to the left (underestimation).

Analyses were not performed on distributions resulting from aggregation with the conjugate model and weights summing to four, the number of individuals in the group. The result of larger weights would be only to decrease considerably the dispersion of the already too tight distributions without changing the accuracy (as measured by the mode).

Figure 6

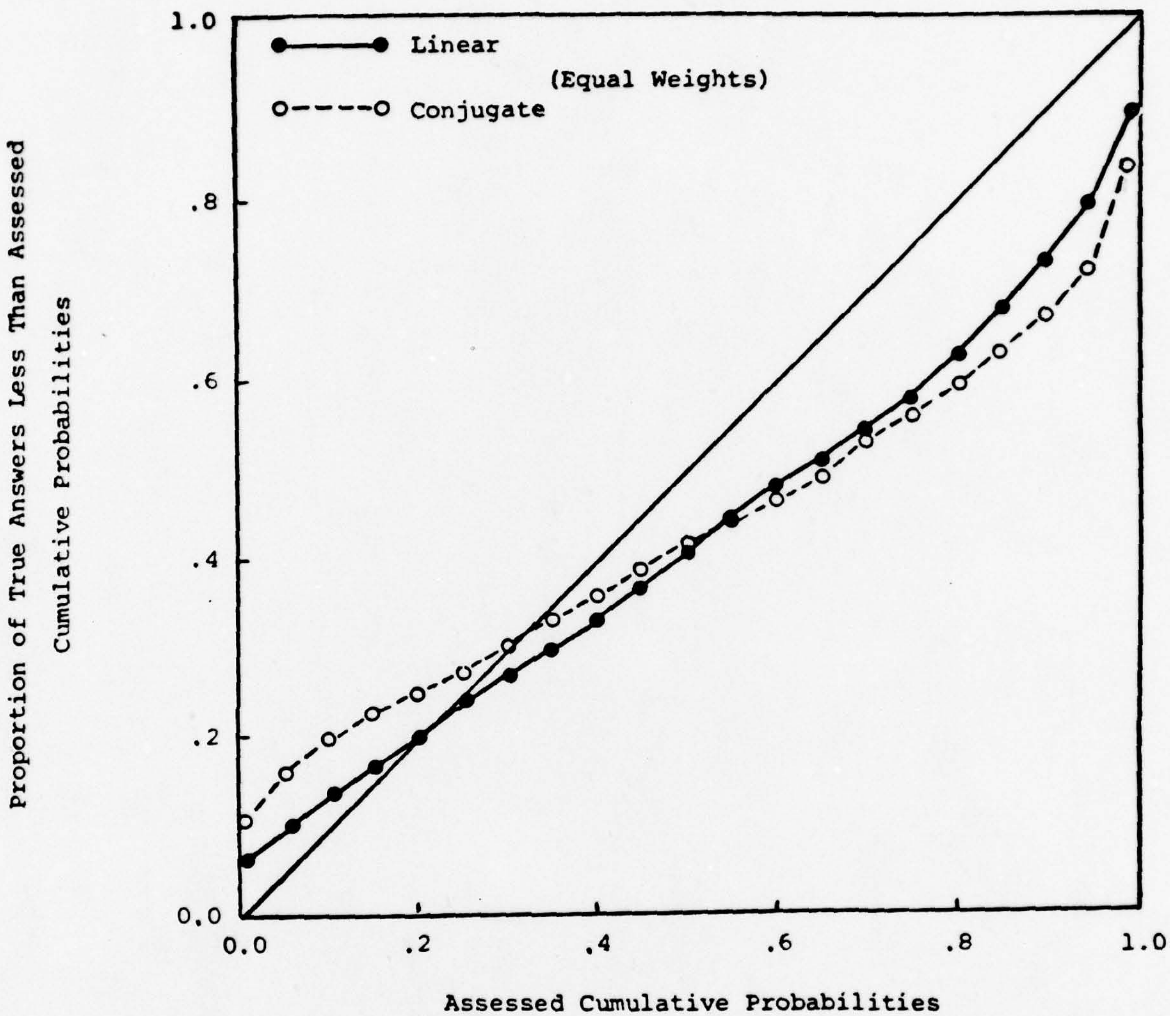
CALIBRATION OF DIFFERENT AGGREGATION MODELS AND WEIGHTING PROCEDURES
BEFORE INTERACTION: CONTINUOUS ASSESSMENTS



Model	Weights	Tails	IQ Range	Maximum Discrepancy
Linear	Equal	5.0%	52.4%	--
Linear	Self-rating	5.3%	50.5%	.027
Linear	DeGroot	5.8%	50.6%	.013
Conjugate	Equal	23.7%	29.8%	--
Conjugate	Self-rating	25.3%	29.4%	.015
Conjugate	DeGroot	25.2%	29.3%	.007

Figure 7

CALIBRATION OF DIFFERENT AGGREGATION MODELS AND WEIGHTING PROCEDURES AFTER INTERACTION: CONTINUOUS ASSESSMENTS



Model	Weights	Tails	IQ Range	Maximum Discrepancy
Linear	Equal	16.1%	34.4%	--
Linear	Self-rating	16.8%	35.0%	.018
Linear	DeGroot	17.3%	34.2%	.011
Conjugate	Equal	27.1%	28.6%	--
Conjugate	Self-rating	28.1%	29.2%	.013
Conjugate	DeGroot	27.9%	28.6%	.010

When questioned at the end of the experiment as to which procedure they would prefer to use in a real decision making situation, subjects exhibited a clear preference for interaction with some open, face-to-face discussion. Twenty subjects prefered the NGT procedure, 19 the CON procedure, and 1 the MIX procedure.

DISCUSSION

As is the case with all studies of the size and complexity of this one, some "significant" results can always be teased out of the data. Rather than focusing on particular significant results, or non-significant ones for that matter, I will discuss some fairly general conclusions and their implications for assessing group probabilities in actual decision making contexts.

The results of this study can be viewed from two perspectives. From the psychological viewpoint, the results are relatively uninteresting. The type of interaction groups are allowed seems to have little effect on subsequent judgments such as those in this study, although all types produce some effects. But the implication for applications of decision theory are important: use simple, mathematical aggregation procedures. Simple procedures, such as combining individual probability assessments linearly with equal weights, produce group assessments that are as good as or better than those produced by more complicated procedures involving interaction or complex aggregation models. Interaction among the assessors produces only a feeling of satisfaction, and not any overall improvement in the quality of the assessed probabilities. Naturally, the results of this study are not as simple and straight-forward as these two viewpoints imply, but they do capture the spirit of this research.

These conclusions are not new or unique to this research. Fischer (1975) concurs with the lack of effect on group probabilities due to the type of interaction, and Gough (1975) presents results that appear to

support this lack of effect, although he does not explicitly adopt such a position. Dalkey (1969b), Gustafson _et al_. (1973), and Van de Ven and Delbecq (1974) have argued in favor of specific interaction procedures; Dalkey for Delphi, and Gustafson _et al_. and Van de Ven and Delbecq for NGT. But Dalkey's conclusion is supported by very weak evidence, and the latter two studies rely on suspect evaluation criteria.

On the model side of the question, the literature indicating little difference in aggregation models due to weighting procedures is becoming extensive (cf. Dawes and Corrigan, 1974; Wainer, 1976). And in contexts other than aggregating probabilities (e.g., multiattribute utility models), linear models have been shown to produce results quite similar to those of non-linear models (Fischer, 1972; Newman, Seaver, and Edwards, 1976). This study has confirmed the lack of effect of different weighting schemes and, at least in the case of discrete assessments, the similarity of results from linear and multiplicative aggregation models for the particular case of aggregating individual probabilities to form a group probability.

The result of interaction among assessors is quite clear for both discrete and continuous assessments--it produces more extreme and less well calibrated assessments. If all of the members of the group agree on an answer, or if even three agree, the individual assessments tend to become more extreme. Apparently, subjects treat the information provided by other group members' assessments as somewhat independent of their own information, rather than redundant. With the particular type of questions and subjects used in this study, this assumption of independence is probably unwarranted, as shown by an analysis

of the data. Since the sum of weights in the multiplicative aggregation model (eq. 5) can be used as an indication of the degree of independence of the individual assessments, the initial individual discrete assessments were aggregated for each group by the multiplicative model with the sum of the weights varying from 1.0 to 4 in steps of .1 and from 4 to 10 in steps of .5. The aggregated assessments were scored with the quadratic scoring rule. The best average score was obtained with the weights summing to 1.0, indicating little independent information in the assessments of different individuals.

Situations where different assessors can be expected to possess somewhat independent information clearly cannot be assumed to produce results similar to those of this study. More extensive modeling may be required in such situations, unless subsequent research shows some type of interaction can be beneficially used. But practical considerations can be used to guide selection of a procedure for determining a group probability when there is no a priori rationale for distinguishing among multiple assessors. Use of a simple mathematical model to aggregate initial individual assessments rather than any type of interaction can lead to considerable savings in time and effort on the part of decision makers or other experts. Linear aggregation is particularly attractive because of its computational simplicity which makes it easily understood, and, therefore, possibly more acceptable. However, simple mathematical aggregation of any sort may not be an acceptable procedure to decision makers. As indicated by the subjects in this study who overwhelmingly preferred some type of interaction with open face-to-face discussion, procedures involving interaction may be desired. If this is true, the

NGT procedure would appear to be the procedure of choice. Although it was generally not "significantly" better than other procedures in this study, it was somewhat better, as it has been in other studies (Gough, 1975; Gustafson et al., 1973).

Snapper and Seaver (1978) provide an example of a situation where mathematical aggregation is a preferable alternative to an interactive process. As part of the evaluation of a national criminal justice program, probabilistic judgments about expected program outcomes are being obtained from experts. Simply averaging these judgments rather than bringing the experts together to interact reduces the logistical complexity and the cost of obtaining the judgments. And as shown by this study, does so with no real loss in the quality of the resulting probability assessments.

REFERENCES

Arrow, K. Social Choice and Individual Values. New York: Wiley, 1951.

Bacharach, M. Group decisions in the face of differences of opinion. Management Science, 1975, 22, 182-191.

Beach, B. Expert judgment about uncertainty: Bayesian decision making in realistic settings. Organizational Behavior and Human Performance, 1975, 14, 10-59.

Brown, T. An experiment in probabilistic forecasting. Rand Report R-944-ARPA, The Rand Corporation, Santa Monica, Ca., 1973.

Collins, B., and Guetzkow, H. A Social Psychology of Group Processes for Decision-Making. New York: Wiley, 1964.

Dalkey, N. Analyses from a group opinion study. Futures, 1969, 1, 541-551. (a)

Dalkey, N. An experimental study of group opinion: The Delphi method. Futures, 1969, 1, 408-426. (b)

Dalkey, N. An impossibility theorem for group probability functions. Rand Paper P-4862, The Rand Corporation, Santa Monica, Ca., 1972.

Dalkey, N. Toward a theory of group estimation. In Linstone, H., and Turoff, M. (Eds.). The Delphi Method: Techniques and Applications. Reading, Ma.: Addison-Wesley, 1975.

Dalkey, N., Brown, B., and Cochran, S. The Delphi method IV: Effect of percentile feedback and feed-in of relevant facts. Rand Memorandum RM-6118-PR, The Rand Corporation, Santa Monica, Ca., 1970. (a)

Dalkey, N., Brown, B., and Cochran, S. Use of self-ratings to improve group estimates. Technological Forecasting, 1970, 1, 283-291. (b)

Dalkey, N., and Helmer, O. An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9, 458-467.

Davis, J. Group Performance. Reading Ma.: Addison-Wesley, 1969.

Dawes, R., and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

DeGroot, M. Optimal Statistical Decisions. New York: McGraw-Hill, 1970.

DeGroot, M. Reaching a consensus. Journal of the American Statistical Association, 1974, 69, 118-121.

Delbecq, A., Van de Ven, A., and Gustafson, D. Group Techniques for Program Planning. Glenview, Il.: Scott, Foresman, 1975.

Eisenberg, E., and Gale, D. Consensus of subjective probabilities: the pari-mutuel method. Annals of Mathematical Statistics, 1959, 30, 165-168.

Fischer, G. Four methods for assessing multi-attribute utilities: An experimental validation. Technical Report 037230-6-T, Engineering Psychology Laboratory, University of Michigan, 1972.

Fischer, G. An experimental study of four procedures for aggregating subjective probability assessments. Technical Report 75-7, Decisions and Designs, Inc., McLean, Va., 1975.

Fujii, T., Seaver, D., and Edwards, W. New and old biases in subjective probability distributions: Do they exist and are they affected by elicitation procedures? SSRI Research Report 77-4, Social Science Research Institute, University of Southern California, 1977.

Goodman, B. Action selection and likelihood ratio estimation by individuals and groups. Organizational Behavior and Human Performance, 1972, 7, 121-141.

Gough, R. The effect of group format on aggregate subjective probability distributions. In Wendt, D., and Vlek, C. (Eds.). Utility, Probability, and Human Decision-Making. Dordrecht-Holland: Reidel, 1975.

Gustafson, D., Shukla, R., Delbecq, A., and Walster, G. A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. Organizational Behavior and Human Performance, 1973, 9, 280-291.

Hogarth, R. Cognitive processes and the assessment of subjective probability distributions. Journal of the American Statistical Association, 1975, 70, 271-294.

Lichtenstein, S., Fischhoff, B., and Phillips, L. Calibration of probabilities: The state of the art. In Jungermann, H., and deZeeuw, G. (Eds.). Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision-Making. Dordecht-Holland: Reidel, 1977.

Linstone, H., and Turoff, M. (Eds.). The Delphi Method: Techniques and Applications. Reading, Ma.: Addison-Wesley, 1975.

Matheson, J., and Winkler, R. Scoring rules for continuous probability distributions. Management Science, 1976, 22, 1087-1096.

Morris, P. Decision analysis expert use. Management Science, 1974, 20, 1233-1241.

Morris, P. Combining expert judgments: A Bayesian approach. Management Science, 1977, 23, 679-693.

Murphy, A., and Winkler, R. Scoring rules in probability assessment and evaluation. Acta Psychologica, 1970, 34, 273-286.

Newman, J. R., Seaver, D., and Edwards, W. Unit versus differential weighting schemes for decision making. SSRI Research Report 76-5, Social Science Research Institute, University of Southern California, 1976.

Norvig, T. Consensus of subjective probabilities: A convergence theorem. Annals of Mathematical Statistics, 1967, 38, 221-225.

Roberts, H. Probabilistic prediction. Journal of the American Statistical Association, 1965, 60, 50-62.

Rowse, G., Gustafson, D., and Ludke, R. Comparison of rules for aggregating subjective likelihood ratios. Organizational Behavior and Human Performance, 1974, 12, 274-285.

Sackman, H. Delphi assessment: Expert opinion, forecasting and group process. Rand Report R-1283-PR, The Rand Corporation, Santa Monica, Ca., 1974.

Savage, L. The Foundations of Statistics. New York: Wiley, 1954.

Seaver, D. Assessment of group preferences and group uncertainty for decision making. SSRI Research Report 76-4, Social Science Research Institute, University of Southern California, 1976.

Seaver, D., von Winterfeldt, D., and Edwards, W. Eliciting subjective probability distributions on continuous variables. Organizational Behavior and Human Performance, 1978, 21, 379-391.

Snapper, K., and Seaver, D. Application of decision analysis to program planning and evaluation. Technical Report 78-1, Decision Science Consortium, Inc., Reston, Va., 1978.

Spetzler, C., and Stael von Holstein, C.-A. Probability encoding in decision analyses. Management Science, 1975, 22, 340-358.

Stael von Holstein, C.-A. Measurement of subjective probability. Acta Psychologica, 1970, 34, 146-159.

Stael von Holstein, C.-A. An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 1971, 10, 635-645.

Stael von Holstein, C.-A. Probabilistic forecasting: An experiment related to the stock market. Organizational Behavior and Human Performance, 1972, 8, 139-158.

Stone, M. The opinion pool. Annals of Mathematical Statistics, 1961, 32, 1339-1342.

Van de Ven, A., and Delbecq, A. Nominal versus interacting group processes for committee decision-making effectiveness. Academy of Management Journal, 1971, 14, 203-212.

Van de Ven, A., and Delbecq, A. The effectiveness of nominal, Delphi, and interacting group decision making processes. Academy of Management Journal, 1974, 17, 605-621.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Winkler, R. The consensus of subjective probability distributions. Management Science, 1968, 15, 61-75.

Winkler, R. Probabilistic prediction: Some experimental results. Journal of the American Statistical Association, 1971, 66, 675-685.

von Winterfeldt, D., and Edwards, W. *Flat maxima in linear optimization* models. Technical Report 011313-4-T, Engineering Psychology Laboratory, University of Michigan, 1973.

CONTRACT DISTRIBUTION LIST
(Unclassified Technical Reports)

Director Advanced Research Projects Agency Attention: Program Management Office 1400 Wilson Boulevard Arlington, Virginia 22209	2 copies
Office of Naval Research Attention: Code 455 800 North Quincy Street Arlington, Virginia 22217	3 copies
Defense Documentation Center Attention: DDC-TC Cameron Station Alexandria, Virginia 22314	12 copies
DCASMA Baltimore Office Attention: Mr. K. Gerasim 300 East Joppa Road Towson, Maryland 21204	1 copy
Director Naval Research Laboratory Attention: Code 2627 Washington, D.C. 20375	6 copies

Revised August 1978

SUPPLEMENTAL DISTRIBUTION LIST
(Unclassified Technical Reports)

Department of Defense

Director of Net Assessment
Office of the Secretary of Defense
Attention: MAJ Robert G. Gough, USAF
The Pentagon, Room 3A930
Washington, DC 20301

Assistant Director (Net Technical Assessment)
Office of the Deputy Director of Defense
Research and Engineering (Test and
Evaluation)
The Pentagon, Room 3C125
Washington, DC 20301

Director, Defense Advanced Research
Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, ARPA Regional Office (Europe)
Headquarters, U.S. European Command
APO New York 09128

Director, ARPA Regional Office (Pacific)
Staff CINCPAC, Box 13
Camp H. M. Smith, Hawaii 96861

Dr. Don Hirta
Defense Systems Management School
Building 202
Ft. Belvoir, VA 22060

Chairman, Department of Curriculum
Development
National War College
Ft. McNair, 4th and P Streets, SW
Washington, DC 20319

Defense Intelligence School
Attention: Professor Douglas E. Hunter
Washington, DC 20374

Vice Director for Production
Management Office (Special Actions)
Defense Intelligence Agency
Room 1E863, The Pentagon
Washington, DC 20301

Command and Control Technical Center
Defense Communications Agency
Attention: Mr. John D. Hwang
Washington, DC 20301

Department of the Navy

Office of the Chief of Naval Operations
(OP-951)
Washington, DC 20450

Office of Naval Research
Assistant Chief for Technology (Code 200
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Director, ONR Branch Office
Attention: Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
Attention: Dr. M. Bertin
American Embassy - Room A-407
APO San Francisco 96503

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps (Code RD-1)
Washington, DC 20380

Headquarters, Naval Material Command
(Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Dean of Research Administration
Naval Postgraduate School
Attention: Patrick C. Parker
Monterey, CA 93940

Superintendent
Naval Postgraduate School
Attention: R. J. Roland, (Code 52RL)
C^3 Curriculum
Monterey, CA 93940

Naval Personnel Research and Development
Center (Code 305)
Attention: LCDR O'Bar
San Diego, CA 92152

Navy Personnel Research and Development
Center
Manned Systems Design (Code 311)
Attention: Dr. Fred Muckler
San Diego, CA 92152

Naval Training Equipment Center
Human Factors Department (Code N215)
Orlando, FL 32813

Naval Training Equipment Center
Training Analysis and Evaluation Group
(Code N-00T)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Director, Center for Advanced Research
Naval War College
Attention: Professor C. Lewis
Newport, RI 02840

Naval Research Laboratory
Communications Sciences Division (Code 54
Attention: Dr. John Shore
Washington, DC 20375

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, MD 21402

Chief, Intelligence Division
Marine Corps Development Center
Quantico, VA 22134

Department of the Army

Deputy Under Secretary of the Army
(Operations Research)
The Pentagon, Room 2E621
Washington, DC 20310

THIS PAGE IS BEST QUALITY PRACTICABLE
FROM COPY FURNISHED TO DDC

Director, Army Library
Army Studies (ASDIRS)
The Pentagon, Room 1A534
Washington, DC 20310

U.S. Army Research Institute
Organizations and Systems Research Laboratory
Attention: Dr. Edgar M. Johnson
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and Systems
Research Laboratory
U.S. Army Institute for the Behavioral
and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Technical Director, U.S. Army Concepts
Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20014

Director, Strategic Studies Institute
U.S. Army Combat Developments Command
Carlisle Barracks, PA 17013

Commandant, Army Logistics Management Center
Attention: DRXMC-LS-SCAD (ORSA)
Ft. Lee, VA 23801

Department of Engineering
United States Military Academy
Attention: COL A. F. Grum
West Point, NY 10996

Marine Corps Representative
U.S. Army War College
Carlisle Barracks, PA 17013

Chief, Studies and Analysis Office
Headquarters, Army Training and Doctrine
Command
Ft. Monroe, VA 23351

Commander, U.S. Army Research Office
(Durham)
Box CM, Duke Station
Durham, NC 27706

Department of the Air Force

Assistant for Requirements Development
and Acquisition Programs
Office of the Deputy Chief of Staff for
Research and Development
The Pentagon, Room 4C331
Washington, DC 20330

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Commandant, Air University
Maxwell AFB, AL 36112

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Deputy Chief of Staff, Plans, and
Operations
Directorate of Concepts (AR/XOCCC)
Attention: Major R. Linhard
The Pentagon, Room 4D 1047
Washington, DC 20330

Director, Advanced Systems Division
(AFHRL/AS)
Attention: Dr. Gordon Eckstrand
Wright-Patterson AFB, OH 45433

Commander, Rome Air Development Center
Attention: Mr. John Atkinson
Griffis AFB
Rome, NY 13440

IRD, Rome Air Development Center
Attention: Mr. Frederic A. Dion
Griffis AFB
Rome, NY 13440

HQS Tactical Air Command
Attention: LTCOL David Dianich
Langley AFB, VA 23665

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, DC 20505

Director, Center for the Study of
Intelligence
Central Intelligence Agency
Attention: Mr. Dean Moor
Washington, DC 20505

Mr. Richard Heuer
Methods & Forecasting Division
Office of Regional and Political Analysis
Central Intelligence Agency
Washington, DC 20505

Office of Life Sciences
Headquarters, National Aeronautics and
Space Administration
Attention: Dr. Stanley Deutsch
600 Independence Avenue
Washington, DC 20546

Other Institutions

Department of Psychology
The Johns Hopkins University
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Stanford University
Attention: Dr. R. A. Howard
Stanford, CA 94305

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A. D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Brunel University
Attention: Dr. Lawrence D. Phillips
Uxbridge, Middlesex UB8 3PH
England

Decision Analysis Group
Stanford Research Institute
Attention: Dr. Miley W. Merkhofer
Menlo Park, CA 94025

Decision Research
1201 Oak Street
Eugene, OR 97401

Department of Psychology
University of Washington
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Department of Electrical and Computer
Engineering
University of Michigan
Attention: Professor Kan Chen
Ann Arbor, MI 94135

Department of Government and Politics
University of Maryland
Attention: Dr. Davis B. Bobrow
College Park, MD 20747

Department of Psychology
Hebrew University
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. Andrew P. Sage
School of Engineering and Applied
Science
University of Virginia
Charlottesville, VA [illegible]

Professor Raymond Tanter
Political Science Department
The University of Michigan
Ann Arbor, MI 48109

Professor Howard Raiffa
Morgan 302
Harvard Business School
Harvard University
Cambridge, MA 02163

Department of Psychology
University of Oklahoma
Attention: Dr. Charles Gettys
455 West Lindsey
Dale Hall Tower
Norman, OK 73069

Institute of Behavioral Science #3
University of Colorado
Attention: Dr. Kenneth Hammond
Room 201
Boulder, Colorado 80309

Decisions and Designs, Incorporated
Suite 600, 8400 Westpark Drive
P.O. Box 907
McLean, Virginia 22101

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 001922-3-T	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): (6) Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation.	5. TYPE OF REPORT & PERIOD COVERED: Technical 10/77 -- 12/78
	6. PERFORMING ORG. REPORT NUMBER: - SSRI 78-3
7. AUTHOR(s): (10) David Arden Seaver	8. CONTRACT OR GRANT NUMBER(s): (15) Prime Contract N00014-76-C-0074; Subcontract 76-030-0715
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Social Science Research Institute, University of Southern California, Los Angeles, CA 90007	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, Virginia 22209	12. REPORT DATE: (11) Dec[redacted]78
	13. NUMBER OF PAGES: 58
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Decisions & Designs, Inc., 8400 Westpark Drive, Suite 600, McLean, Virginia 22101 (under contract from Office of Naval Research) (12) 66p.	15. SECURITY CLASS. (of this report): unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

(9) Research rept. Oct 77-Dec 78

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THIS PAGE IS BEST QUALITY PRACTICABLE FROM COPY FURNISHED TO DDC

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

decision analysis
group decision making
subjective probability
group judgment
Delphi method
nominal group
statisticized group

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The application of decison theory often involves assessing subjective probabilities and procedures for assessing them are quite well developed. But such procedures are based on assessments by a single person. Often multiple individuals are called on to provide the probabilistic judgments. Unanimity in judgments among the multiple individuals cannot be expected, thereby creating the problem of how to arrive at a single probability distribution that can be used in applying decision theory. Two general approaches to this problem exist. The individuals

over

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

390 664 Du

can interact as a group to reach a consensus, or the individual jugments can be mathematically aggregated to produce a single probability distribution. Each of these approaches has advantages and disadvantages. Group interaction allows the exchange of information, but may be susceptible to dominance by certain individuals or pressure for conformity. Mathematical aggregation is simple to use and ensures that a single distribution will result, but theoretical difficulties are encountered in specifying an appropriate aggregation model. Using several forms of group interaction and mathematical aggregation models, this research investigated the quality of probabilities produced by interaction versus mathematical models. "Quality" was measured by proper scoring rules, calibration, and extremeness on two types of probability assessments: discrete assessments for two-alternative questions and beta probability density functions for questions about percentages. Ten four-person groups comprised primarily of graduate students assessed probabilities for twenty questions of each type in each of five types of group interaction: no interaction, Delphi, Nominal Group Technique (NGT) a mix of Delphi and NGT, and discussion to consensus. The mathematical models used to aggregate the individual assessments included linear model, the weighted geometric mean, and the pari-mutuel model for discrete assessments; and the linear model and conjugate model for densities; each with various weighting procedures. Applying proper scoring rules to the group probabilities indicated that simple mathematical aggregation without any interaction, e.g. linear aggregation with equal weights, generally produced group probabilities as good as those assessed after interaction. Interaction did produce more extreme but less well calibrated assessments, with the type of interaction having little effect. Generally, the calibration of mathematically aggregated group probabilities prior to any interaction was quite good, clearly better than the calibration of individual assessments. These results may appear relatively uninteresting from a psychological perspective because of the lack of differences in assessments after different types of interaction. But the implications for applications of decision theory are important. In many instances, simple, mathematical aggregation of individual probability assessments may be adequate without resorting to more elaborate, practically difficult, and time consuming interactive processes or modeling efforts.